**UNIVERSIDAD AUTÓNOMA DE MADRID**
**ESCUELA POLITÉCNICA SUPERIOR**

Máster Universitario en Bioinformática y Biología Computacional

TRABAJO FIN DE MÁSTER

# Diversity and dynamics of Escherichia coli ST131 causing bacteremia in a tertiary hospital in Madrid (1996-2016) using genomic tools and latest generation bioinformatics

**Autor: Claire Jordan Brooks**
**Tutores: Teresa Coque González y Val Fernandez Lanza**
**Ponente: Daniel Aguirre De Carcer Garcia**

**Julio 2019**

# Diversity and dynamics of Escherichia coli ST131 causing bacteremia in a tertiary hospital in Madrid (1996-2016) using genomic tools and latest generation bioinformatics

Autor: Claire Jordan Brooks
Tutores: Teresa Coque González y Val Fernandez Lanza
Ponente: Daniel Aguirre De Carcer Garcia

Instituto Ramón y Cajal de Investigación Sanitaria

Escuela Politécnica Superior Universidad
Autónoma de Madri
Julio 2019

# Abstract

**Motivation:** Extraintestinal pathogenic *E. coli* is the leading cause of urinary tract infections and of adult bacteremia. The clonal group STc131 has been identified as the primary cause of the global pandemic of multidrug resistance within these *E. coli*.

**Objective**: To analyze the diversity of STc131 from a representative sample of isolates of phylogroup B2 E. coli that caused bacteremia in the hospital Ramón y Cajal in the last 20 years.

**Material and Methods**: Eighty-three E. coli strains, primarily identified as STc131 from a representative sample of 528 B2-EC isolates associated with non-duplicated episodes of bacteremia in Hospital Ramón y Cajal from the years 1996-2016, were selected for next generation sequencing. Bioinformatic techniques were used to annotate the genomes and to identify epidemiological markers and mobile genetic elements that may contribute to the pathogenic success of particular STc131 clades.

**Results**: The number of clade C subgroups of STc131 samples isolated from episodes of bacteremia in Hospital Ramón y Cajal has increased over the period of collection (1996-2016), while clade B maintained a stable frequency. Analysis of the accessory genomes of these isolates indicate a high degree of diversity between and within the different STc131 clades. Less than a quarter (23.7%) of the sample were ESBL positive.

**Conclusions**. These data indicate the importance of the accessory genome, including the plasmid content of the isolates, in the dynamics of this STc131 population with each clade having a distinct plasmid, virulence, and antibiotic resistance signature. The results highlight diversification of the STc131 before and after the selection by first line antibiotics including fluoroquinolones and cephalosporins.

# Key Words

Antibiotic Resistance, Bacterial Genomics, Comparative Genomics, Phylogenetic Analysis

# Acknowledgements

This study was supported by the Joint Programming Initiative in Antimicrobial Resistance (ST131, JPIonAMR2016-AC16/00039), the Instituto de Salud Carlos III of Spain/Ministry of Economy and Competitiveness- and the European Development Regional Fund "A way to achieve Europe" (ERDF) for co-founding the Spanish R&D National Plan Estatal de I+D+i 2013-2016 (PI15-0512, P18-1942 ), CIBER (CIBER in Epidemiology and Public Health, CIBERESP; CB06/02/0053), and the Regional Government of Madrid (InGeMICS- B2017/BMD-3691).

I would like to thank my advisors Teresa Coque González, Val Fernández Lanza and Daniel Aguirre de Carcer García for all their guidance and support during the realization of this TFM. Their insights have been invaluable, and their advice was key in my professional development.

# General Index

# Figure Index



# Table Index

# 1

# Introduction

## 1.1. Motivation of the project

Extra-intestinal pathogenic *Escherichia coli* (ExPEC) are *E. coli* able to cause infections outside the gastrointestinal system and are the leading cause of urinary tract infections and of adult bacteremia[1,2]. While ExPEC was mostly susceptible to the first line antibiotics that are used to treat these infections, such as cephalosporins (cephs) and fluoroquinolones (FQs) before the 2000s[2], FQs have become increasingly ineffective in patients throughout the world[3]. The clonal group STc131 has been identified as the primary cause of the global pandemic of multidrug resistance of ExPEC[4,5]. Up to 30% of all ExPEC, 60-90% of FQ-R ExPEC and 40-80% of extended spectrum beta lactamase (ESBL) ExPEC belongs to STc131[6].

The STc131 high-risk clonal complex is subdivided into different genetic groups according to the allelic profile of fimH, and the profiles of resistance and virulence (A, B, C0, C1, C2)[7–10]. Clade C2, which is highly associated with the FQ-R and ESBL production of STc131, has been globally dominant since the year 2000[4]. Recent papers indicate that the origin of Clade C STc131 correlates to the introduction of FQ treatments in the mid-to-late 1980s, propelled in large part by chromosomal mutations within the gyrA and parC genes that confer FQ-R[11,12]. However, these factors cannot fully explain the global dominance clade C has achieved, as other ExPEC strains with similar genetic profiles have not adapted to the increasing use of FQ and cephs[12] as quickly or as effectively[13].

## 1.2 Objectives and Focus

It is likely that mobile genetic elements, such as plasmids and conjugative transposons, have contributed to the success of STc131. However the exact mechanisms by which these factors contributed to that success remain unclear. Previous studies have shown a clear association between clade type and plasmid incompatibility groups[14] Notably, plasmids associated with C2 (FII:A1:B replicons) carry beneficial ESBL and virulence factors, such as *bla*$_{\text{CTX-M-15}}$ and the pemI/pemK addiction system[12,15]. When the integration of these plasmids occurred and whether it contributed to the diversification of clade C into C1 and C2 has not yet been elucidated.

Developing a detailed characterization of clade C STc131 provides a unique opportunity to clarify the processes leading to high-risk transmission of antimicrobial resistance and virulence in emerging MDR clones. However, the over-representation of sequenced genomes of the C2 subgroups and the scarcity of data on the STc131 accessory genome have made it difficult to reconstruct the evolutionary history of this polyclonal complex.

This project aims to analyze the diversity of STc131 from a representative sample of all isolates of E. coli STc131 that caused bacteremia in the hospital Ramón y Cajal in the last 20 years.

## 1.3. Methodology and Work Plan

In this project NGS technology was used to investigate the roles of MGEs on the fitness of STc131 in order to delineate the evolutionary history of this high-risk clone. Bioinformatic techniques were used to annotate the genome and to identify genomic islands that may contribute to the success of STc131 clade C[11]. AccNET, an application based on bipartite networks developed in our lab, was used to assess the "evolvability" of STc131 clades through the analysis of the accessory genome[16]. Analysis of plasmids to determine relationship between plasmid types and STc131 subclades, was performed with Hyasp, Mob-Suite, PlasmidFinder, and the CGE pMLST database. Comparison of the Ramón y Cajal STc131 samples were compared against NCBI draft genomes of STc131 and non-STc131 phylogroup B2 *E. coli*.

# 2

# *E. coli* STc131, multi-drug resistance, and the role of accessory genome on pathogenic success

## 2.1. Introduction

*Escherichia coli* is categorized into eight recognized phylogroups: A, B1, B2, C, D, E, F, and clade I[17]. Phylogroups A and B1 are most associated with commensal strains, while phylogroups B2 and D are associated with ExPEC strains[18–20]. Phylogroup B2 can be divided into 9 subgroups, where the basal subgroup B2-I is demonstrated to have high virulence and diversity, and contains the fewest B2 flexible gene pool sequences (set of sequences derived from comparative analysis of pathogenic B2 strains to commensal phylogroup A strains)[21]. The clonal group STc131 is a phylogroup B2-I ExPEC strain with a high incidence of MDR, shown to express greater resistance and virulence profiles than non-STc131 phylogroup B2 MDR strains[6,22–24]. While STc131 has a robust virulence profile, notably certain classical B2 virulence genes are present at low frequency or absent in STc131 populations (P fimbria *pap* genes, cytotoxic necrotizing factor *cnf1*, haemolysin A *hlyA*, increased serum survival gene *iss*, and siderophore receptor *iroN*)[6,9,25].

## 2.2. STc131 Clade Diversification

### 2.2.1 Differentiating the three clades of STc131 (A, B, and C)

The three primary clades of the clonal group STc131 can be resolved by FimH typing, core-genome multi-locus strain typing (cgMLST), and the presence of FQ-R conferring mutations within DNA gyrase subunit A (gyrA) and topoisomerase IV (parC)[14,26,27].

The Type 1 fimbrial adhesin FimH is widespread in *E. coli* and ubiquitous in STc131[9]. The diversity of the *FimH* gene has emerged as a valuable typing schema within E. coli groups[28,29]. Within STc131, the fimH41 is associated with clade A, fimH22 with clade B, and fimH30 with clade C[30]. Serotyping additionally distinguishes clade A (serotype O16:H5) from clades B and C (serotype O25b:H4)[13,31]. The Pasteur Institute's cgMLST scheme is another typing method that has been shown to be effective at discriminating between clonal lineages and can be used to identify subgroups within STc131[32–34]. In this schema, clade A is associated with pST506, clade B with pST9, and clade C with pST43 and pST621[14,35].

### 2.2.2. Characteristics of STc131 Clade C

Gyrase subunit A is a common target of quinolones and mutations within the base range 67-106 of gyrA, known as the quinolone-resistance determining region (QRDR), have been

demonstrated to confer quinolone resistance[36,37]. Fluoroquinolone resistance conferring point mutations within gyrA (D87N, known as allele gyrA1AB) and parC (S80I and E84V, known as allele parC1aAB) further differentiate clade C from the rest of STc131[27,30]. Clade C (fimH30) STc131 with these mutations are known as H30-R, while a smaller subgroup of fimH30 STc131 that lack this chromosomal fluoroquinolone resistance are known as non-H30-R, or clade C0[13,38]. The existence of clade C0 and genetic homogeneity within H30-R support a stepwise evolution model of diversification, wherein H30-R evolved from a single strain of fimH30 STc131 that acquired the specific combination of gyrA/parC mutations that confer FQ-R[38]. However, there have also been cases of non-fimH30 STc131 strains with gyrA/parC FQ-R mutations[27].

Clade C is discriminated into two subclades: H30-R (subclade C1) and H30-Rx (subclade C2), where C2 is noted by the association with the *bla*$_{\text{CTX-M-15}}$ gene[23,27,31,39,40]. While CTX-M-15 is primarily found in C2, it is neither ubiquitous or exclusive to this subclade. The Cefotaximase-Munich[41] (CTX-M-) type ESBLs are a group of class A β-lactamases with over a hundred variants and the ability to hydrolyze cefotaxime and other extended-spectrum cephalosporin (ESC) antimicrobials[42,43]. While CTX-M-15 is the predominant CTX-M-type within STc131[44], CTX-M-14 has been associated to a lesser extent with non-H30 clades and CTX-M-27 has been associated with H30-R subclade C1[25,38,45]. It has been recently suggested that subclade C1 should be further subcategorized into C1 and C1-M27, to represent the emergent *bla*$_{\text{CTX-M-27}}$ positive cluster that has been identified in Japan[46].

### 2.2.3 Timeline of the Evolution of STc131

Longitudinal NGS analyses of historical STc13 isolates demonstrate that clade C0 (acquisition of fimH30 by recombination, acquisition of genomic islands GI-*leuX* and GI-*pheV*) arose in North America from clade B in the late 1970s/early 1980s. This was followed by the development of clade C (acquisition of gyrA/parC FQ-R mutations) in the mid/late 1980s, which quickly divided into subclades C1 and C2[11,12]. Thus the emergence of FQ-R STc131 coincides with the introduction of second generation quinolone (fluoroquinolone) antimicrobials[47]. It is unknown when subclade C2 acquired the now characteristic *bla*$_{\text{CTX-M-27}}$ gene, though other studies show that STc131 clade C did not become the dominant *E. coli* lineage until a rapid expansion during the 2000s[8,27] and that subclade C2 did not become dominant until the late 2000s, indicating that the acquisition of the ESBL occurred sometime in this time period[40].

## 2.3. Success of STc131 Clade C

The global increase in ESBL producing bacteria over the past decade has been widely attributed to STc131 and more recently attributed to subclade C2[6,25,27,35,48–52]. It has been suggested that the acquisition of genomic islands such as the virulence-rich GI-*pheV* (which carries several virulence genes including *sat* and *iuta*), along with the acquisition of FQ-R chromosomal mutations, had primed subclade C2 for success upon acquisition of *bla*$_{\text{CTX-M-15}}$[11,38]. This is consistent with evidence that STc131 is an exception to the typical "trade-off" between virulence and antibiotic resistance previously observed at the phylogroup

level in *E. coli*[53], as *bla*$_{CTX-M-15}$ positive STc131 strains have been shown to have high colonization ability and lethality in mice models[54,55] despite lacking many of the genes associated with the virulence potential of phylogroup B2[56,57]. Later studies show extreme variability in the lethality of both FQ-R and FQ-S STc131 strains, indicating a high degree of virulence diversity within the clonal group and individual clades[57,58].

## 2.4. Accessory Genome and Mobile Genetic Elements

### 2.4.1 Genomic Islands

While the acquisition of genomic islands such as GI-*pheV* have been identified as key events in the differentiation of clade C from the rest of STc131[11], this clonal group lacks many of the pathogenicity islands (PAIs) associated with other ExPEC strains[55]. Of these four PAIs (PAI I$_{CFT073}$, PAI II$_{J96}$, PAI III$_{536}$, and the Yersinia high-pathogenicity island [HPI]), only HPI is highly associated with STc131. However, HPI is not uniquely characteristic to STc131 and occurs frequently throughout phylogroup B2 *E. coli*, studies showing that more than 90% of B2 strains contain markers for this island[59,60]. Subgroups of both clades B (pST9) and C (pST621) have also been identified that are positive for *hyla* and *cnf1*, markers for PAI II$_{J96}$, however the PAI II$_{J96}$ positive strains paradoxically present less lethality in mice models than similar virotypes without this PAI[61,58]. This indicates that a specific combination, or balance, of accessory virulence traits has greater influence on the virulence potential of STc131 than the number of virulence genes alone.

### 2.4.2 Plasmidome

The acquisition of different plasmids has been shown to be a key aspect in the variability between the resistomes and virulence of STc131 clades and subclades, most notably in the expression of *bla*$_{CTX-M-15}$ in C2[12,13,62,63]. The incompatibility group F (IncF) family of plasmids occur throughout STc131 clades, make up the majority of plasmid types identified within STc131 samples, and are associated with the dissemination of ESBL genes within the clonal group[12,14,64], while non-F incompatibility groups (N, I1/K/BO, I2, A/C, X) carrying antibiotic resistance traits are rare in *E. coli*[65]. IncF plasmids are a heterogeneous group which encompasses a range of sizes (90-140 kb) and replicon types, including repFII, repFIA, and repFIB (FAB), which may or may not occur on the same plasmid[14,19]. The allelic variants within the FAB replicons are a common typing scheme for plasmids of this type[66].

subclade C1 is typically associated with non-*bla*$_{CTX-M-15}$ plasmids with the IncF replicon sequence type (RST) FI:A2:B20, while C2 is associated with *bla*$_{CTX-M-15}$ positive plasmids with either the IncF RST FII:A1:B-[63] or F31:A4:B1/F36:A4:B1[67]. However, it has been shown that *bla*$_{CTX-M-15}$ is carried within STc131 on a variety of IncF plasmids with a variety of replicon sequence types[68,69]. In subclade C2, the *bla*$_{CTX-M-15}$ gene has been shown to consistently co-occur with the *blaTEM-1*, *bla*$_{OXA-1}$, the *aac(6')-Ib-cr* genes, suggesting a common point of origin[19]

# 3

# Materials and Methods

*All software and versions outlined in this section are noted in the Software Appendix at the end of this document.*

## 3.1 Phylogroup B2 *E. coli* Sample collection and typing

Eighty-three E. coli (EC) strains, primarily identified as STc131 from a sample of 528 B2-EC[70] isolates representative of 5870 strains associated with non-duplicated episodes of bacteremia from the years 1996-2016, were selected for next generation sequencing (NGS) with illumina.

## 3.2 DNA sequencing, read processing, genome assembly, and quality control

Total DNA was extracted from 5 mL of overnight cultures using the Wizard Genomic DNA Purification Kit (Promega Corp., Madison, WI, USA) and the DNA concentration was measured using a Qubit™ Fluorometer and Nanodrop 1000 (Thermo Scientific, Waltham, MA, USA). The library preparation was carried out with the Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA). Sequencing was performed using a standard 2 x 100 or 2 x 150 base protocol in a Genome Analyzer IIx Illumina HiSeq 2500 platform (Illumina, San Diego, CA).

The reads were initially processed with Trimmomatic (slidingwindow:5:25 leading:10 trailing:10 minlen:40) and evaluated for quality with FastQC. The main read statistics of the 83 sequence datasets analyzed are shown in Table 1. The trimmed reads were assembled using the SPAdes short reads assembler on *careful* mode using kmer sizes of 27, 33, 39, 45, 51, and 55. Assembly quality was assessed with Quast and one sample was removed (total assembly size smaller than expected of an E. coli genome at 44kb, with N50 and L50 orders of magnitude smaller than the rest of the sample).

## 3.3 Identification of *E. coli* sequence types and clade diversity

Genome annotation was performed with Prokka against the *Escherichia* protein BLAST database. Strain typing was performed using the Achtman scheme of multi-locus strain typing (MLST) and core genome scheme of multi-locus strain typing (cgMLST). This was accomplished through screening of the PubMLST database using Torsten Seeman's MLST software. Non-STc131 assemblies were excluded from further processing. For STc131 strains, this was followed by fimH-typing (CGE's fimtyper) and identification of *gyrA/parC* chromosomal mutations (CGE's pointfinder) for clade and subclade identification.

**Table 1. Read statistics for 83 phylogroup B2 illumina sequences[1]**

| Assembly | Reads | Avg Len | GC% | Avg Phred | Coverage | Assembly | Reads | Avg Len | GC% | Avg Phred | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 1528655 | 125 | 50.2 | 39.8 | 36x | **43** | 2487245 | 125 | 50.0 | 39.8 | 59x |
| **2** | 1900498 | 125 | 50.3 | 39.8 | 45x | **44** | 1949680 | 125 | 50.1 | 39.7 | 46x |
| **3** | 2348636 | 125 | 50.3 | 39.8 | 56x | **45** | 2492071 | 125 | 50.2 | 39.7 | 59x |
| **4** | 2340073 | 125 | 49.9 | 39.7 | 55x | **46** | 2429223 | 125 | 50.0 | 39.7 | 57x |
| **5** | 2042788 | 125 | 50.3 | 39.7 | 48x | **47** | 2085165 | 125 | 49.9 | 39.8 | 49x |
| **6** | 2814352 | 125 | 50.4 | 39.7 | 67x | **48** | 1176902 | 126 | 50.1 | 39.9 | 28x |
| **7** | 1850088 | 125 | 50.1 | 39.9 | 44x | **49** | 1509640 | 125 | 50.1 | 39.8 | 36x |
| **8** | 1588117 | 125 | 49.9 | 39.9 | 38x | **50** | 1746141 | 126 | 50.1 | 39.9 | 41x |
| **9** | 1438409 | 126 | 50.1 | 39.8 | 34x | **51** | 1681436 | 125 | 50.1 | 39.8 | 40x |
| **10** | 1561687 | 125 | 50.1 | 39.8 | 37x | **52** | 2344382 | 125 | 50.1 | 39.8 | 56x |
| **11** | 2719195 | 125 | 50.3 | 39.8 | 64x | **53** | 2601471 | 125 | 50.2 | 39.7 | 61x |
| **12** | 2256231 | 125 | 50.4 | 39.8 | 53x | **54** | 2674833 | 125 | 50.5 | 39.8 | 63x |
| **13** | 2550827 | 125 | 50.3 | 39.7 | 60x | **55** | 2459848 | 125 | 49.9 | 39.7 | 58x |
| **14** | 2279946 | 125 | 50.2 | 39.8 | 54x | **56** | 2397001 | 125 | 50.2 | 39.9 | 57x |
| **15** | 1653931 | 125 | 50.0 | 39.8 | 39x | **57** | 1982524 | 125 | 50.1 | 39.8 | 47x |
| **16** | 1890486 | 125 | 49.7 | 39.8 | 45x | **58** | 33477296 | 79 | 49.0 | 37.8 | 507x |
| **17** | 1891847 | 125 | 50.2 | 39.8 | 45x | **59** | 40094098 | 79 | 49.1 | 37.8 | 607x |
| **18** | 1400808 | 125 | 49.9 | 39.8 | 33x | **60** | 15074148 | 79 | 49.0 | 37.8 | 228x |
| **19** | 2278049 | 125 | 50.3 | 39.8 | 54x | **61** | 15816514 | 79 | 48.8 | 37.7 | 239x |
| **20** | 2582971 | 125 | 50.3 | 39.7 | 61x | **62** | 41125814 | 79 | 48.9 | 37.9 | 623x |
| **21** | 2209882 | 125 | 49.9 | 39.7 | 52x | **63** | 43619906 | 79 | 49.2 | 37.9 | 661x |
| **22** | 2086333 | 125 | 50.1 | 39.8 | 49x | **64** | 45225972 | 79 | 48.5 | 37.5 | 681x |
| **23** | 2498043 | 125 | 50.0 | 39.8 | 59x | **65** | 20285068 | 78 | 48.6 | 37.4 | 304x |
| **24** | 2649158 | 125 | 50.1 | 39.7 | 63x | **66** | 20224694 | 78 | 48.7 | 37.4 | 303x |
| **25** | 2214284 | 125 | 50.1 | 39.8 | 52x | **67** | 17562682 | 78 | 48.9 | 37.4 | 263x |
| **26** | 2374179 | 125 | 50.3 | 39.8 | 56x | **68** | 49730322 | 79 | 48.5 | 37.6 | 750x |
| **27** | 1717471 | 125 | 50.1 | 39.8 | 41x | **69** | 37932706 | 79 | 48.5 | 37.5 | 570x |
| **28** | 1769159 | 125 | 50.1 | 39.7 | 42x | **70** | 22633554 | 78 | 48.9 | 37.4 | 339x |
| **29** | 1627535 | 125 | 50.1 | 39.7 | 38x | **71** | 52945406 | 79 | 49.0 | 37.5 | 797x |
| **30** | 2484841 | 125 | 49.9 | 39.8 | 59x | **72** | 25651630 | 79 | 48.9 | 37.4 | 385x |
| **31** | 1849420 | 125 | 50.0 | 39.8 | 44x | **73** | 19510332 | 78 | 49.0 | 37.3 | 292x |
| **32** | 2360689 | 125 | 50.1 | 39.8 | 56x | **74** | 38629076 | 79 | 48.5 | 37.5 | 581x |
| **33** | 2594986 | 125 | 50.6 | 39.8 | 62x | **75** | 23320540 | 78 | 48.6 | 37.4 | 350x |
| **34** | 3076073 | 125 | 50.8 | 39.7 | 73x | **76** | 22379164 | 79 | 49.0 | 37.5 | 336x |
| **35** | 2643006 | 125 | 50.0 | 39.8 | 63x | **77** | 27322788 | 79 | 48.5 | 37.5 | 411x |
| **36** | 2379647 | 125 | 50.4 | 39.7 | 56x | **78** | 29414934 | 79 | 48.7 | 37.6 | 442x |
| **37** | 3025544 | 125 | 50.6 | 39.8 | 72x | **79** | 26487640 | 79 | 48.6 | 37.6 | 398x |
| **38** | 2773984 | 125 | 50.2 | 39.8 | 66x | **80** | 25752898 | 79 | 48.7 | 37.6 | 387x |
| **39** | 121567 | 125 | 50.1 | 39.7 | ***2x*** | **81** | 11890930 | 78 | 48.6 | 37.4 | 178x |
| **40** | 1377721 | 125 | 50.3 | 39.7 | 32x | **82** | 13356202 | 78 | 50.1 | 37.2 | 198x |
| **41** | 2718951 | 125 | 49.9 | 39.8 | 64x | **83** | 13325888 | 78 | 49.3 | 37.2 | 198x |
| **42** | 2021300 | 125 | 50.1 | 39.7 | 48x | | | | | | |

[1]isolates 1-57 sequenced in 2018 using 2 x 150 protocol, isolates 58-83 processed in 2014 using 2 x 100 protocol

## 3.4. Core, Accessory and Whole genome analysis

The alignment and identification of SNPs was performed using Snippy rapid haploid variant calling and core genome alignment of the original reads against an O25b-STc131 representative genome from Ensembl (IEH71520_1.0), recombination events were filtered using Gubbins, and the construction of the core genome maximum likelihood phylogenetic tree using IQtree with 100 bootstrap replicates.

Comparison of core, whole, and accessory genome analyses was performed to evaluate the diversification of STc131 clades and subclades at different genomic strata. The 73 STc131 strains used in this analysis were pooled with an additional 25 STc131 isolates (including members of clade A and clade C1 positive for *bla*$_{CTX\text{-}M\text{-}27}$) that were also collected at Hospital Ramón y Cajal. The accessory genome was analyzed with Accnet, delineating the set core and accessory genome products. The accnet data was additionally processed with Eggnog Mapper to functionally annotate the accessory genome based on Clusters of Orthologous Groups of proteins (COGs), using all orthologs and non-electronic terms to

prioritize coverage. The whole genome was analyzed by mash distance, producing a distance matrix by Mash pairwise distance calculation on Mash genome sketches with a k-mer of 27. The SNP core and accessory genome information was converted into distance matrices determined by the number of shared traits (common SNPs for core alignment, common accessory genes for accessory genome analysis) between every pair of assemblies. The resulting distance matrices from the SNP core, whole, and accessory genomes were clustered with Mclust and principal component analysis and visualized was performed with fviz. We performed hypergeometric testing to determine significant proteins in each resulting cluster, adjusting the P-value by the Benjamini-Hochberg procedure to account for multiple comparisons.

The construction of the accessory and whole genome phylogenetic trees were constructed by neighbor joining using the ape R package. All phylogenetic trees were visualized using the iTOL v4. Interactive Tree Of Life web interface.

## 3.5. Epidemiological Markers

The screening (identity 90%, coverage 90%) antibiotic resistance and virulence genes was performed using Abricate against the Resfinder databases available at the Center for Genome Epidemiology (CGE, www.cge.cbs.dtu.dk) and a repository of known *E. coli* virulence genes (from VFDB, http://www.mgc.ac.cn/VFs/) supplemented with factors supported by literature from the Canadian National Microbiology Laboratory (https://github.com/phac-nml/ecoli_vf). Visualization of identified pathogenicity islands was performed using genoPlotR[71].

## 3.6. Plasmidome analysis

Plasmid replicon screening was performed against the PlasmidFinder database with abricate software (90% coverage, 90% identity), followed by replicon sequence typing on IncF plasmids using the pMLST incF RST database. MOB relaxase and oriT typing was performed with MOB-Suite software (90% coverage, 90% identity).

Multiple plasmid reconstruction methods were evaluated, including the MOB-suite plasmid reconstruction and Hyasp assembler packages. MOB-suite was used for plasmid assembly and typing using the MOB-recon method, which utilizes similarity to clusters of complete plasmids to reconstruct plasmids from whole genome data. The HyASP hybrid assembler utilizes similarity to discrete plasmid-associated genes (90% identity, 90% coverage threshold against the Hyasp database of dereplicated genes of the plasmids obtained from NCBI) to identify potential plasmid contigs, followed by the use of factors such as GC% to aggregate contigs into plasmid groups (minimum gene density: 0.3, minimum length: 1500 bases, maximum GC content variability: 0.15). The resulting sets of MOB-suite and Hyasp plasmid data were then re-screened for incF replicon sequence type to determine any loss of information from these methods. The Hyasp produced plasmid set presented worse recall than the MOB-suite plasmid set, with a 20% loss of FAB replicon data between the whole genome data and proposed plasmid contigs. Thus, MOB-suite was primarily utilized for IncF plasmid reconstruction.

NuRig was used for alignment and visualization of the representative plasmids.

## 3.7. NCBI STc131 Dataset Creation and Analysis

A collection of 797 STc131 draft genomes were curated from NCBI. The *E. coli* assemblies from NCBI were first screened by Mash distance to a O25b-STc131 representative genome from Ensembl (IEH71520_1.0) using the distance cut-off of 0.01 (a number observed in our dataset of B2 E. coli to represent a natural division between sequence types). The resulting set of assemblies were then verified by MLST typing. The dataset underwent a second screening of Mash distance to remove clonal samples, using a distance cut-off of 0.0001 to remove redundant genomes. Phylogenetic tree construction was performed by neighbor joining with Phylip and visualized using Itol. The resulting dataset underwent accessory genome analysis with accnet, screening for virulence and antibiotic resistance genes with Abricate against the resfinder and VFDB databases, and IncF plasmid replicon typed by pMLST, by the parameters previously described in this section.

## 3.8. NCBI Non-STc131 Phylogroup B2 Dataset Creation and Analysis

A collection of 1709 non-STc131 phylogroup B2 draft genomes were curated from NCBI. Hierarchal cluster analysis was performed with hclust on the Mash whole genome pairwise distance matrix of the *E. coli* assemblies from NCBI. A range of k values (5-200, step of 5) were tested for the optimal number of clusters to capture phylogroup B2 samples, selecting a value of k=75. The 797 STc131 assemblies were filtered out of the cluster and the remaining set of assemblies were then verified by MLST typing. The highest frequency MLSTs within this dataset were of known phylogroup B2 sequence types (323 ST95, 264 ST73, 115 ST13, 92 ST127, 42 ST1193, 30 ST141), while the remainder of the dataset was comprised of many low-frequency variants. The NCBI phylogroup B2 dataset then underwent a second screening of Mash distance to remove clonal samples, using a distance cut-off of 0.0001 to remove redundant genomes. The resulting dataset underwent screening for virulence and antibiotic resistance genes with Abricate against the resfinder and VFDB databases using previously described parameters.

# 4

# Results

## 4.1 STc131 Clade Differentiation

The 73 STc131 samples were grouped into three main branches corresponding to clade A (fimH41 n = 1), clade B (fimH22 n = 13, fimH324 n = 1, fim388 n = 1) and clade C (fimH30 n = 57, fimH35 n = 1, subgroups C0, C1, C2). Less than a quarter (23.7%) of the strains were producers of ESBL (blaCTX-M-15, -14,-27) (Figure 1.A). Two of the clade C samples are subcategorized as clade C0 (without the characteristic gyrA and parC chromosomal mutations which confer FQ-resistance), 26 were subcategorized as subclade C1 (clade containing isolates with *bla*$_{\text{CTX-M-15}}$) and 29 were subcategorized as subclade C2 (clade without *bla*$_{\text{CTX-M-15}}$ containing isolates).

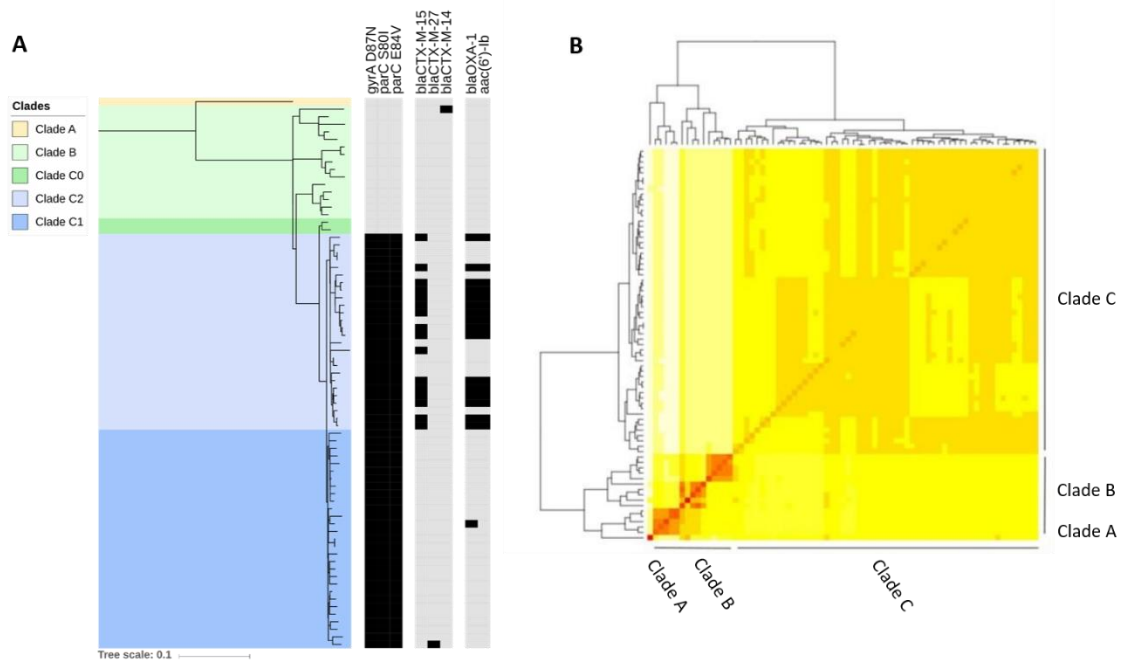


**Figure 1. A)** Maximum likelihood phylogenetic tree of the 2449 SNPS of the core genome of 73 STc131 samples collected between 1996-2017. FimH type (H41,H22,H30, and minor subtypes), presence of 3 FQ-R chromosomal mutations (gyrA D87N, parC S80I and E84V) and ESBL and closely related genes (blaOXA-1, aac(6')-lb) used to categorize samples into different subclades: A,B,C0, C1, and C2. **B)** SNP heat map comparing three different STc131 clades. Approximately 300 SNP distance between clade A and the other two clades, 300 SNP distance between clades B and C, 74 SNP distance between clade C0 and the rest of clade C, and 30 SNP distance between subclades C1 and C2.

## 4.2 Core, Whole, and Accessory Genome Analysis

Performing clustering with Mclust on the SNP core genome of the 98 pooled STc131 isolates (73 from the longitudinal bacteremia study, with 25 additional samples collected from Hospital Ramón y Cajal) demonstrates that the assemblies fall within clusters consistent with each of the three main STc131 clusters (A, B, and C), where the greatest variation occurs between clade A and the other two clades (Figure 2). Within the whole genome cluster analysis there is more separation within clades, as clades B and C sort into three additional clusters each. Within the whole genome analysis, the greatest variation between subclusters occurs within clade B. The accessory genome analysis also sorts clades B and C sort into three separate clusters each. The accessory genome analysis captures the most variance between subgroups of clade C, where clade C1 comprises one cluster and clade C2 separates into two additional groups. In each of these analyses, clade C0 is grouped with either clade C (Core genome) or subclade C1 (accessory and whole genome). The *bla*$_{CTX\text{-}M\text{-}27}$ positive subgroup within clade C was not resolved in either whole or accessory genome analysis.

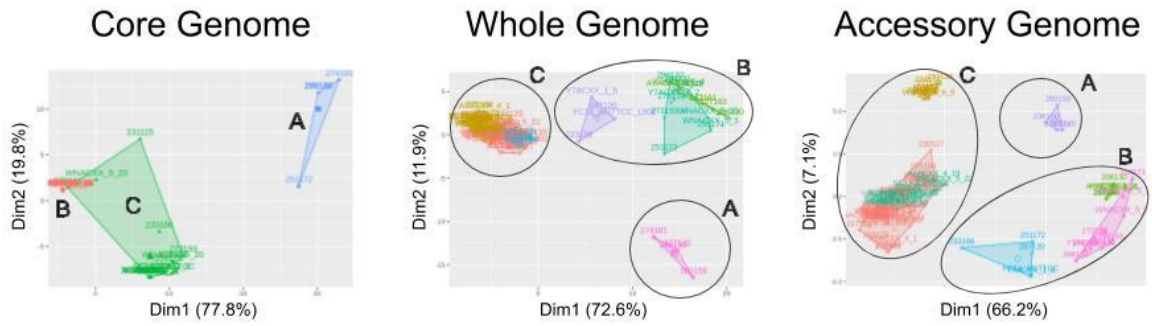


**Figure 2.** Principal Component Analysis representation of core genome, whole genome, and accessory genome STc131 clusters resulting from gaussian mixture model clustering on the with Mclust on the SNP alignment, Mash, and accent distance matrices respectively..

Of the 8446 accessory genome products identified by ACCNET, 1090 were found to be significantly associated (P<0.01) to particular accessory genome clusters, of which 374 were non-hypothetical products. 59 of these non-hypothetical products were associated with clade A clusters, 148 associated with clade B clusters, and 208 to clade C clusters (bearing in mind that some products were significantly associated to multiple clusters). The significant genome products of the clusters comprising each clade were functionally analyzed with Eggnogger (Figure 3.A), as well as the three individual clusters of clade C (Figure 3.B). Overall the functional distribution of the accessory genome followed a similar pattern regardless of clade, with some exceptions. Notably, within both the clade and sub-clade analysis, the group considered the most commensal (clade A and sub-clade C1 respectively) has a strong bias towards proteins associated with carbohydrate transport and

metabolism (COG: G) and against proteins of unknown function (COG: S), compared to the other groups.

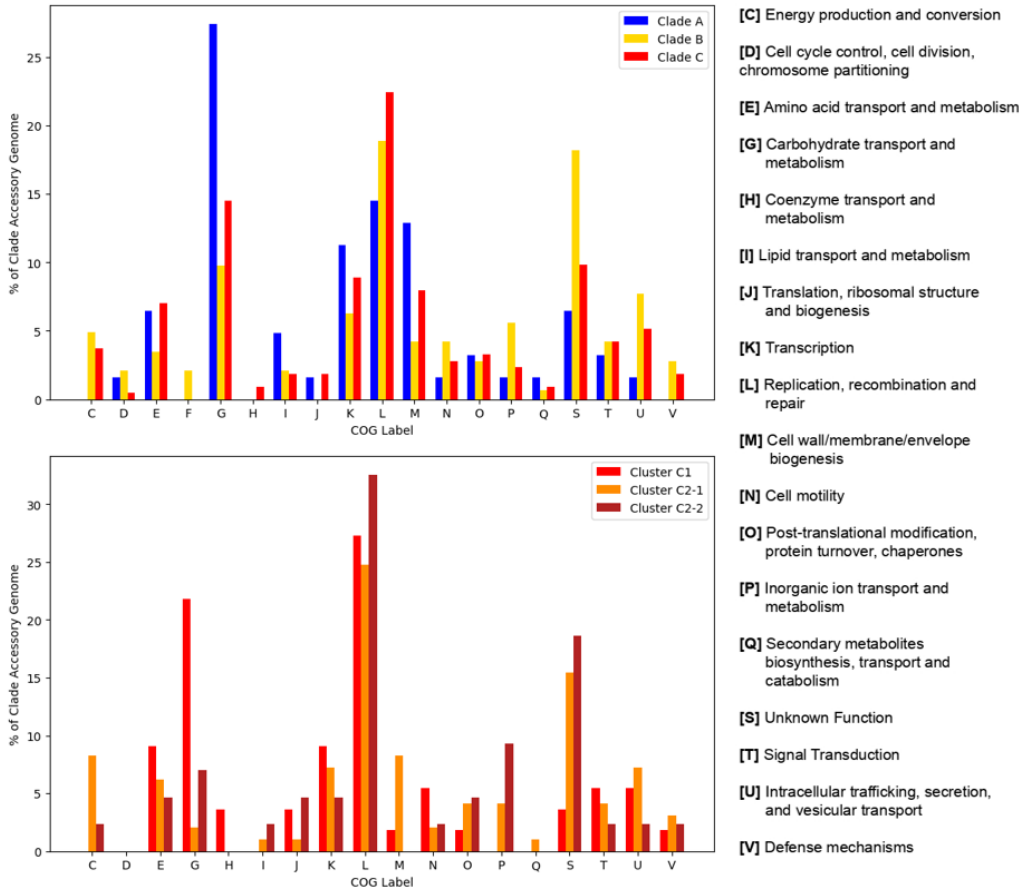


**Figure 3. A)** Percentage of clade associated accessory genome products of each COG functional group. **B)** Percentage of clade C associated accessory genome products of each COG functional group by cluster group determined by accnet (here expressed related to their corresponding subclades: one cluster within C1, and two clusters within C2).

As the accessory genome analysis was found to best represent the diversity within clade types, the accnet analysis and clustering was repeated on the 73 longitudinal samples alone. With this sample 6 clusters were identified: subclade C1 corresponds to cluster one, subclade C2 to clusters three and four, clade B to clusters four, five, and six, while the singular clade A isolate also sorted into cluster six. Of the 6867 accessory genome products identified by accnet, 879 were determined to be significantly associated to particular accessory genome clusters, of which 282 were non-hypothetical products.

## 4.3 STc131 Epidemiological Markers

### 4.3.1 Virulence Profile

A set of virulence genes/groups (*afaA*, *cdtB*, *cnf1*, *hlyA*, *ibeA*, *int*, *iroN*, *tia*, and *papC*) were determined to be significantly associated with the accessory genome clusters (Table 2). These genes closely correspond to those previously determined to be markers of different STc131 virotypes[9,58]. Using the Blanco nine virotype schema, our sample contains: 10 virotype A isolates, 35 virotype C, 2 virotype D1, 3 virotype D2, 4 virotype D3, 4 virotype D5, 8 virotype E. Seven isolates displayed combinations of traits that are not accounted for by this schema. There were no virotype B isolates in our sample.

The virotype distribution closely aligns with the accessory genome clusters: cluster 1 associated with virotype C, cluster 2 with virotype A, cluster 3 with virotype E, cluster 4 with virotypes D1/D2, cluster 5 with virotype D5, and cluster 6 with virotype D3. To reflect the grouping observed by accessory genome cluster and virotype within clade C, we recategorize cluster 2 of subclade C2 into subclade C2-1 (virotype A) and cluster 3 of subclade C2 to subclade C2-2 (virotype E). The hylA/cnf1 fragment of the pathogenicity island PAI $II_{J96}$ that differentiates virotypes E and D4 from similar virulence profiles was identified in a virotype E isolate (Figure 5).

**Table 2. Significant virulence genes by accessory genome cluster**

| Cluster | Gene | Cluster Frequency | Total Frequency | Adjusted P-Value |
|---|---|---|---|---|
| 2 | afaA | 63.60% | 12.00% | 4.04E-04 |
| 2 | afaB-I | 63.60% | 12.00% | 4.04E-04 |
| 2 | afaC-I | 63.60% | 12.00% | 4.04E-04 |
| 2 | afaD | 63.60% | 12.00% | 4.04E-04 |
| 2 | daaF | 63.60% | 12.00% | 4.04E-04 |
| 2 | draE2 | 63.60% | 10.70% | 4.04E-04 |
| 2 | draP | 63.60% | 12.00% | 4.04E-04 |
| 3 | cnf1 | 88.90% | 17.30% | 3.74E-05 |
| 3 | hlyA | 100.00% | 18.70% | 1.46E-06 |
| 3 | hlyB | 100.00% | 18.70% | 1.46E-06 |
| 3 | hlyC | 100.00% | 18.70% | 1.46E-06 |
| 3 | hlyD | 100.00% | 18.70% | 1.46E-06 |
| 3 | papC | 100.00% | 24.00% | 2.95E-05 |
| 3 | papD | 100.00% | 28.00% | 8.23E-05 |
| 3 | papE | 100.00% | 25.30% | 3.74E-05 |
| 3 | papF | 100.00% | 26.70% | 5.10E-05 |
| 3 | papG | 100.00% | 14.70% | 2.00E-07 |
| 3 | papH | 100.00% | 26.70% | 5.10E-05 |
| 3 | papJ | 100.00% | 25.30% | 3.74E-05 |
| 3 | papK | 100.00% | 26.70% | 5.10E-05 |

| | | | | |
|---|---|---|---|---|
| 4 | ibeA | 100.00% | 17.30% | 4.42E-03 |
| 5 | cnf1 | 100.00% | 17.30% | 5.01E-03 |
| 5 | hlyA | 100.00% | 18.70% | 5.19E-03 |
| 5 | hlyB | 100.00% | 18.70% | 5.19E-03 |
| 5 | hlyC | 100.00% | 18.70% | 5.19E-03 |
| 5 | hlyD | 100.00% | 18.70% | 5.19E-03 |
| 5 | ibeA | 100.00% | 17.30% | 5.01E-03 |
| 5 | iroB | 100.00% | 12.00% | 6.54E-04 |
| 5 | iroC | 100.00% | 12.00% | 6.54E-04 |
| 5 | iroD | 100.00% | 12.00% | 6.54E-04 |
| 5 | iroE | 100.00% | 12.00% | 6.54E-04 |
| 5 | iroN | 100.00% | 12.00% | 6.54E-04 |
| 5 | papA | 100.00% | 6.70% | 3.11E-05 |

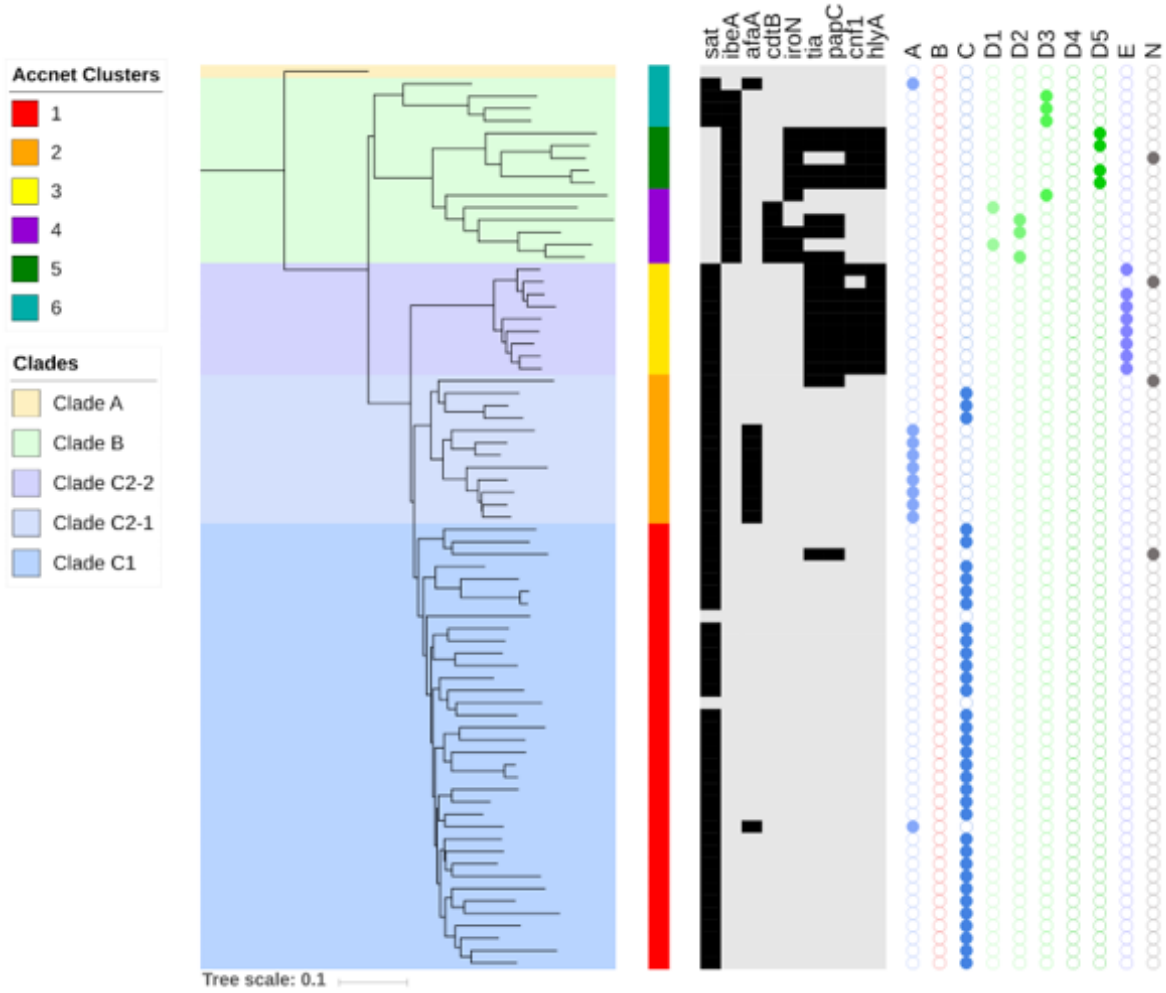


**Figure 4**. Virulence-related genes significantly associated to accessory genome clusters and corresponding virotype (A-E) of 73 Ramón y Cajal STc131 samples. Novel virotypes denoted as "N". Phylogenetic tree constructed by neighbor-joining method on accnet (accessory genome) distance matrix.

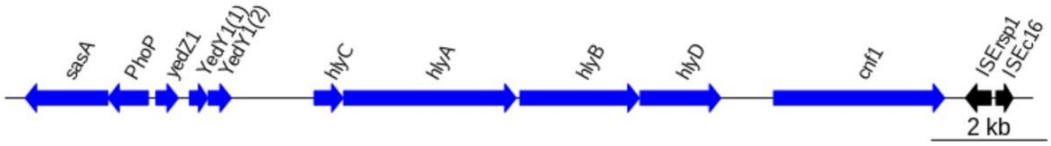


**Figure 5.** Pathogenicity Island (PAI) PIIJ96 fragment from virotype E isolate

4.3.2. Antibiotic Resistance Genes

Nearly a quarter of the 73 STc131 isolates contained ESBL genes (n = 18). The majority of the ESBL genes identified were *bla*$_{\text{CTX-M-15}}$ (16 isolates), of which all but one were situated within subclade C2. One *bla*$_{\textit{CTX-M-27}}$ was identified within subclade C1, and one within *bla*$_{\textit{CTX-M-14}}$ within clade B.

The STc131 clades and subclades additionally have distinct non-ESBL antibiotic resistance signatures, identified by those ABR genes determined to be significantly associated to accessory genome clusters (Table 3). The resistome of clade C is shown to be more extensive than that of clades A and B, containing ABR traits in addition to the ESBL genes concentrated in this clade (Figure 6). The antibiotic resistance profiles of the three accessory genomes within clade C are distinct but share many similarities. The resistome of subclades C1 and C2-1 are nearly identical, apart from the *bla*$_{\text{CTX-M-15}}$ and the related genes (*bla*$_{\text{OXA-1}}$, *aac(61)-lb*) found in C2-1. The resistance profile of subclade C2-2 is also highly similar to that of C2-1, but lacks the *bla*$_{\text{TEM-1B}}$ gene and instead contains the *aac(3)-IIa* gene. We observe that the *bla*$_{\text{TEM-1C}}$ gene is exclusive to clade B.

**Table 3. Significant ABR genes by accessory genome cluster**

| Cluster | Gene | Cluster Frequency | Total Frequency | Adjusted P-Value |
|---|---|---|---|---|
| 1 | aac(3)-IId_1 | 27.80% | 9.60% | 3.68E-04 |
| 1 | aadA5_1 | 55.60% | 28.10% | 2.44E-04 |
| 1 | aph(3")-Ib_5 | 33.30% | 12.30% | 2.44E-04 |
| 1 | aph(6)-Id_1 | 30.60% | 12.30% | 1.43E-03 |
| 1 | blaTEM-1B_1 | 75.00% | 33.30% | 2.43E-08 |
| 1 | dfrA17_1 | 55.60% | 28.10% | 2.44E-04 |
| 1 | mdf(A)_1 | 100.00% | 64.00% | 5.87E-08 |
| 1 | mph(A)_1 | 52.80% | 28.90% | 1.56E-03 |
| 1 | mph(A)_2 | 52.80% | 28.90% | 1.56E-03 |
| 1 | sul1_5 | 58.30% | 31.60% | 4.84E-04 |
| 1 | sul2_2 | 33.30% | 11.40% | 6.24E-05 |
| 2 | aac(6')-Ib-cr_1 | 66.70% | 14.00% | 3.14E-04 |
| 2 | blaCTX-M-15_1 | 66.70% | 14.00% | 3.14E-04 |
| 2 | blaOXA-1_1 | 66.70% | 14.90% | 3.82E-04 |
| 2 | mph(A)_1 | 75.00% | 28.90% | 9.40E-03 |
| 2 | mph(A)_2 | 75.00% | 28.90% | 9.40E-03 |

| | | | | |
|---|---|---|---|---|
| 3 | aac(3)-IIa_1 | 88.90% | 8.80% | 2.40E-08 |
| 3 | aac(6')-Ib-cr_1 | 88.90% | 14.00% | 3.26E-06 |
| 3 | blaCTX-M-15_1 | 77.80% | 14.00% | 7.13E-05 |
| 3 | blaOXA-1_1 | 88.90% | 14.90% | 4.07E-06 |
| 5 | blaTEM-1C_1 | 100.00% | 7.00% | 6.92E-06 |

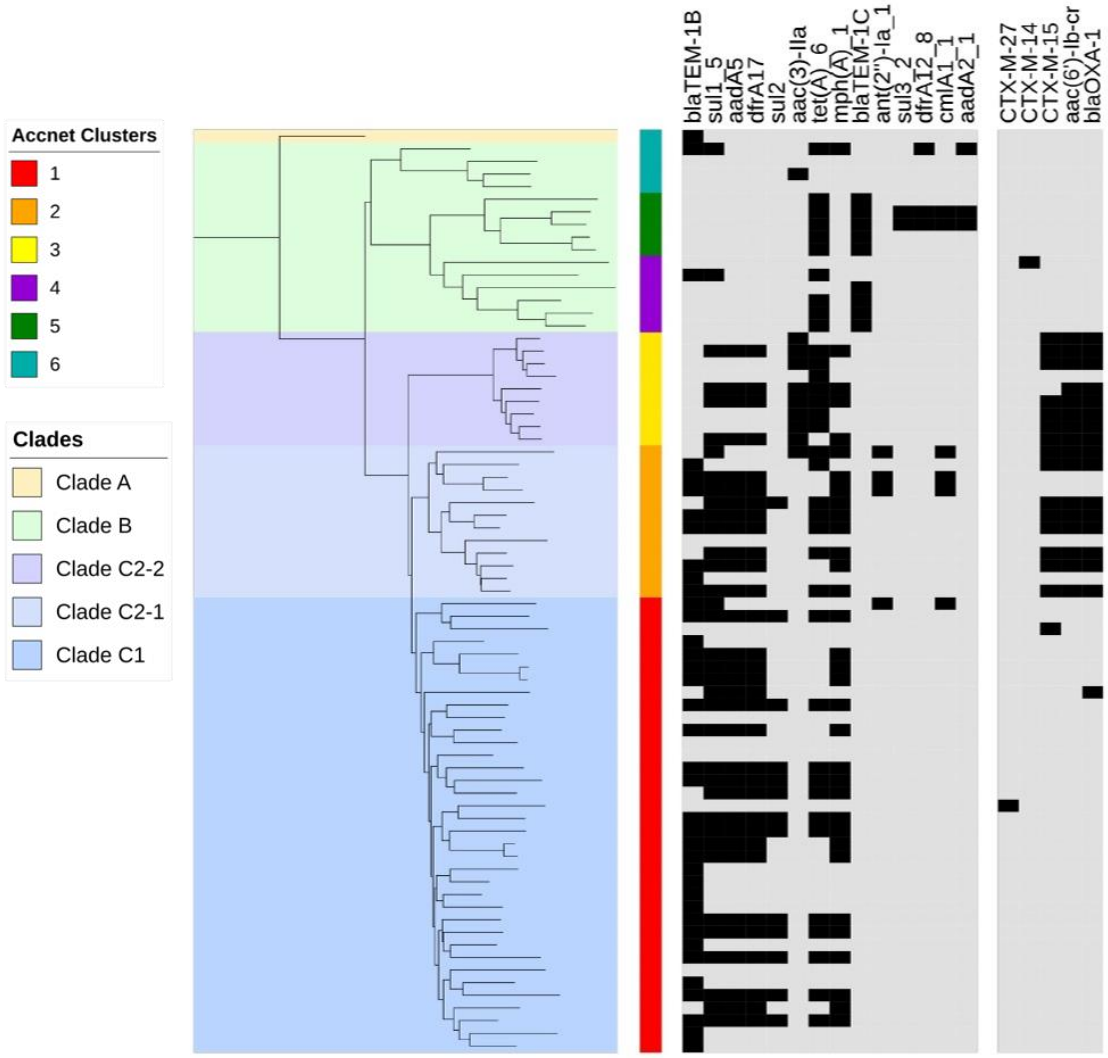


**Figure 6.** Antibiotic resistance genes determined to be significantly associated to accessory genome clusters of 73 Ramón y Cajal STc131 samples. Phylogenetic tree constructed by neighbor-joining method on accnet (accessory genome) distance matrix.

## 4.4 STc131 Plasmidome Analysis

### 4.4.1 IncF Plasmid Replicon Typing

Isolates were screened for incF replicons against the plasmidfinder database, then for specific incF FAB allele type against the CGE pMLST database (Figure 7). The sample contained 12 distinct FAB allele combinations, associated with specific clades and sub-clades of STc131.

The single clade A isolate contained replicon type of F29:A-:B10. Clade B contained eight isolates with replicon type F1:A-:B1, three replicon type F29:A-:B10, one F2:A1:B-. Four clade B lacked an FII allele. Subclade C1 contained 19 isolates of replicon type F1:A2:B20, five F1:A2:B-, four F4:A2:B20, four F2:A1:B-, one F36:A1:B20, and one F2:A-:B-. Subclade C2-1 contained 11 isolates of replicon F2:A1:B-, and one F-:A2:B-. Subclade C2-2 was the most homogeneous, containing 4 isolates of type F36:A4:B1 and 5 of type F31:A4:B1 (a single point mutation exists between the F31 and F36 alleles).

Subclade C2-2 is the only group of STc131 that showed no plasmid permeability with other STc131 clades. There are isolates within both clade A and clade B which contain plasmids with F29:A-:B10 replicon sequence types and isolates within clade B and subclades C1 and C2-1 which contain plasmids with the replicon sequence type F2:A1:B-. Subclade C1 displays the highest IncF replicon variability within the STc131 sample, containing five unique RSTs. In two cases the difference between replicon type were isolated to the presence/absence of a single FAB replicon (F1:A2:B20 and F1:A2:B-, F2:A1:B- and F2:A-:B-).

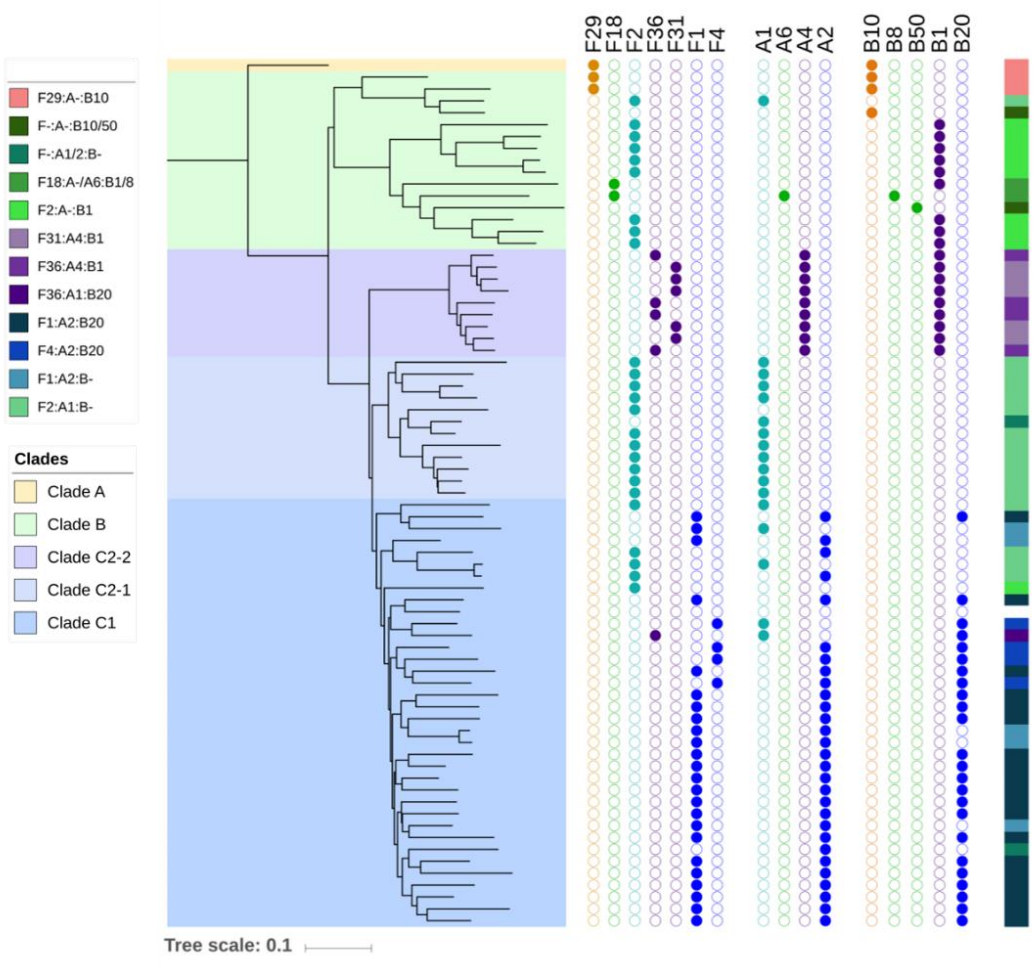


**Figure 7.** Allele type for each incF FAB replicon and composite RST for 73 Ramón y Cajal STc131 samples. Phylogenetic tree constructed by neighbor-joining method on accnet (accessory genome) distance matrix.

### 4.4.2 IncF/MobF Plasmid Extraction

Of the 15 most frequent MOB-suite plasmid clusters identified within our sample (Figure 8), we identified the extracted plasmid with the smallest Mash distance to a reference NCBI plasmid in that cluster to serve as representative of that plasmid cluster within our sample. Seven of these representative plasmids contained both IncF replicon and MobF relaxase types (Table 2). Of these IncF/MOBF plasmids, the sizes range from 69.6kb to 138 kb in size. We see overlap between 1552 cluster plasmids, which occur in both subclades C1 and C2-1. Alignment of these plasmids shows that the C1-1552 plasmid lacks the characteristic $bla_{CTX\text{-}M\text{-}15}$ operon of the 1552 plasmid found in subclade C2-1 (Figure 9). Consistent with what was observed in the whole genome virulence and antibiotic resistance screening, plasmids associated with clade B (973, 1563) have a more robust virulence signature than antibiotic resistance signature, and those associated with clade C have a more robust antibiotic resistance signature.

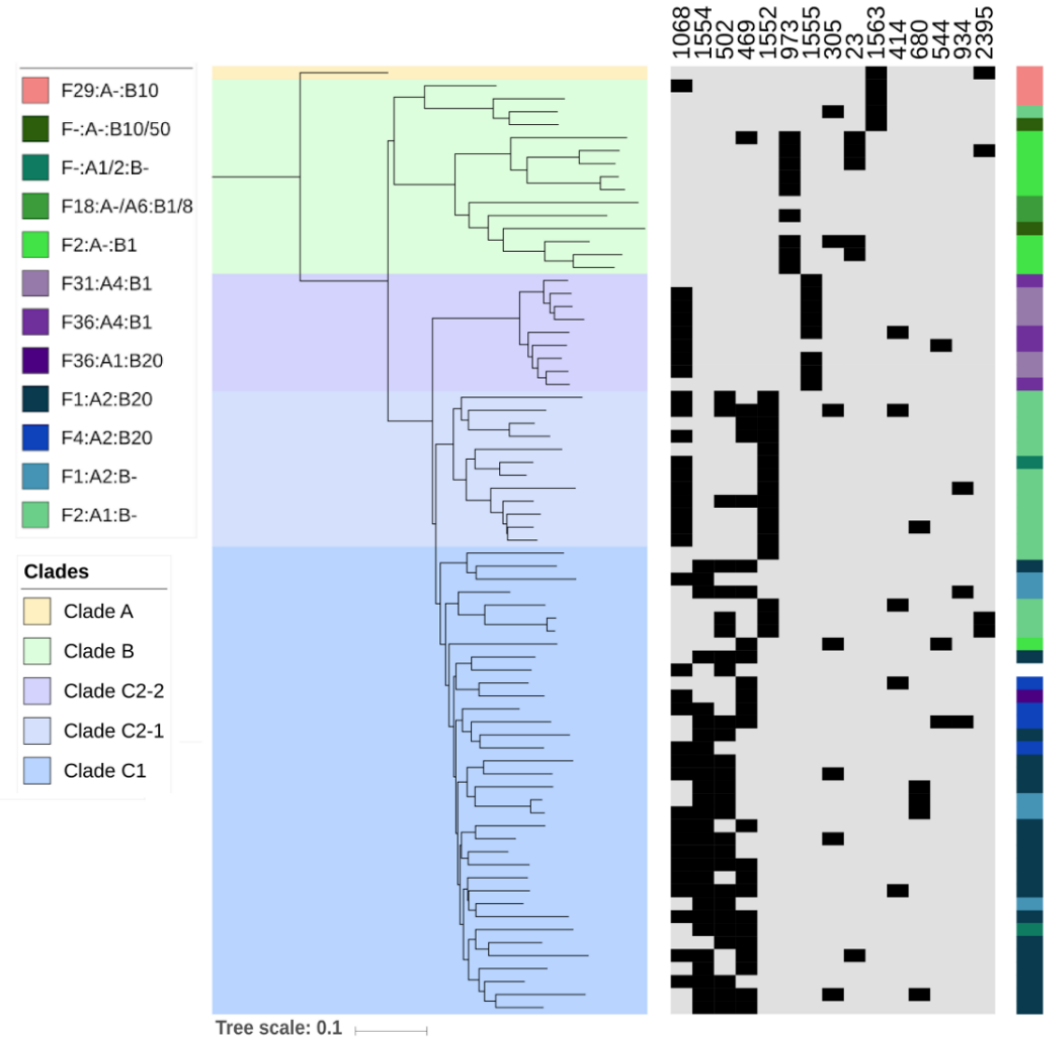


**Figure 8.** Distribution of the 15 most frequent MOB-suite plasmid clusters in comparison to the IncF replicon pMLST distribution. Phylogenetic tree constructed by neighbor-joining method on accnet (accessory genome) distance matrix

**Table 4. Resistance genes and virulence determinants in MOBF plasmids.**

| Cluster | Reference Plasmid | Size (kb) | Rep | Antibiotic Resistance Genes | Virulence Genes |
|---|---|---|---|---|---|
| 544 | KX008967 | 69.6 | IncFIIA (F4:A-:B-) | blaTEM-1B,aadA5, dfrA17, mph(A), sul1, aph(3")-Ib | None-detected |
| 973 | KY007017 | 134 | IncFIB,IncFIIA (F2:A-:B1) | None-detected | iroB,iroC,IroD,iroE ,iroN,iucA,iucB,iuc C, iucD, iutA |
| C1-1552 | CP018978 | 109 | IncFIA,IncFIIA (F2:A1:B-) | blaTEM-1B,aadA5, dfrA17, mph(A), sul1 | None-detected |
| C2-1552 | CP013657 | 138 | IncFIA,IncFIIA (F2:A1:B-) | blaTEM-1B,aadA5, dfrA17, mph(A), sul1,aac(6')-Ib-cr, **blaCTX-M-15**, blaOXA-1 | None-detected |
| 1554 | CP015070 | 106 | IncFIA,IncFIB,IncFIIA (F1:A2:B20) | blaTEM-1B,aadA5, dfrA17, mph(A), sul1 | senB |
| 1555 | KP789020 | 102 | IncFIA,IncFIB,IncFIIA (F31:A4:B1) | aac(6')-Ib-cr, **blaCTX-M-15**, blaOXA-1,aac(3)-IIa_1 | iutA |
| 1563 | CP012634 | 110 | IncFIB,IncFIIA (F29:A-:B10) | None-detected | senB |

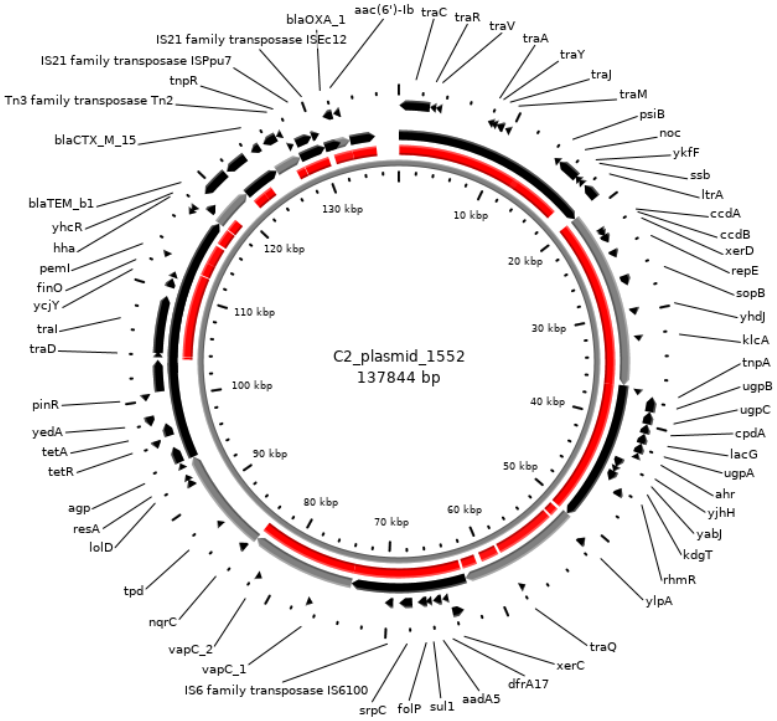


**Figure 9.** C2_Plasmid_1552 (scaffold, black) alignment with C1_plasmid_1552 (red), excluding hypothetical protein data.

## 4.5 Hospital Ramón y Cajal STc131 Longitudinal Analysis

The total number of STc131 isolates identified from the 528 B2-EC strains associated with episodes of bacteremia in Hospital Ramon y Cajal has increased relative to the collection year (Figure 10). The number of clade B isolates identified has remained consistent over the period sampled, however they make up a decreasing proportion of the total STc131 sample as the population of clade C increased, comprising 75.0% of the sample between the years 1996-2004, 17.2% between the years 2005-2010, and 6.45% between the years 2011-2016. The earliest subclade C1 identified was collected in the year 2000, the earliest subclade C2-1 isolate was collected in the year 2005, and the earliest subclade C2-2 isolate three years later in 2008. subclade C1 has been the dominant subgroup of STc131 since the year 2005 (48.3% of the sample 2005-2010 and 60.6% of the sample 2011-2016).

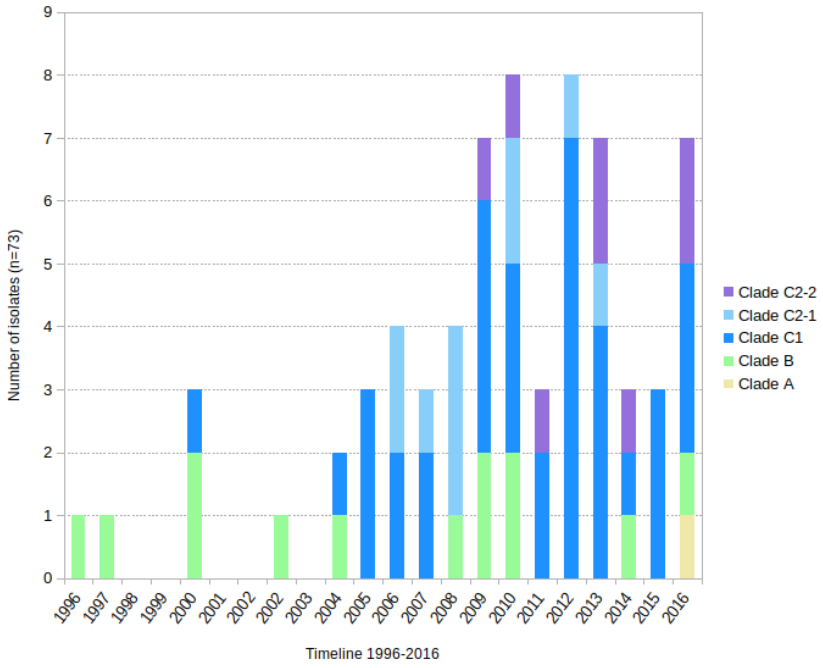


**Figure 10.** Clade distribution of strains collected between the years of 1996 and 2016

## 4.6 NCBI STc131

### 4.6.1 NCBI STc131 Accessory Genome Analysis

The accessory genome delineated by ACCNET analysis of the 797 STc131 assemblies curated from the NCBI database grouped into nine distinct clusters, compared to the seven clusters of the 98 Hospital Ramón y Cajal sample. In the NCBI dataset, Clade A sorted into one ACCNET cluster, clade B sorted into two clusters, subclade C1 sorted into 3 main clusters, and C2 sorted into two clusters. One cluster (cluster 2) spans both subclades C1 and C2, though is primarily associated with C2 (71.0% of the cluster is associated with this subclade). Unlike the Ramón y Cajal sample, the *bla*$_{\text{CTX-M-15}}$ subgroup of subclade C1 was resolved by the accessory genome cluster analysis (Figure 11).

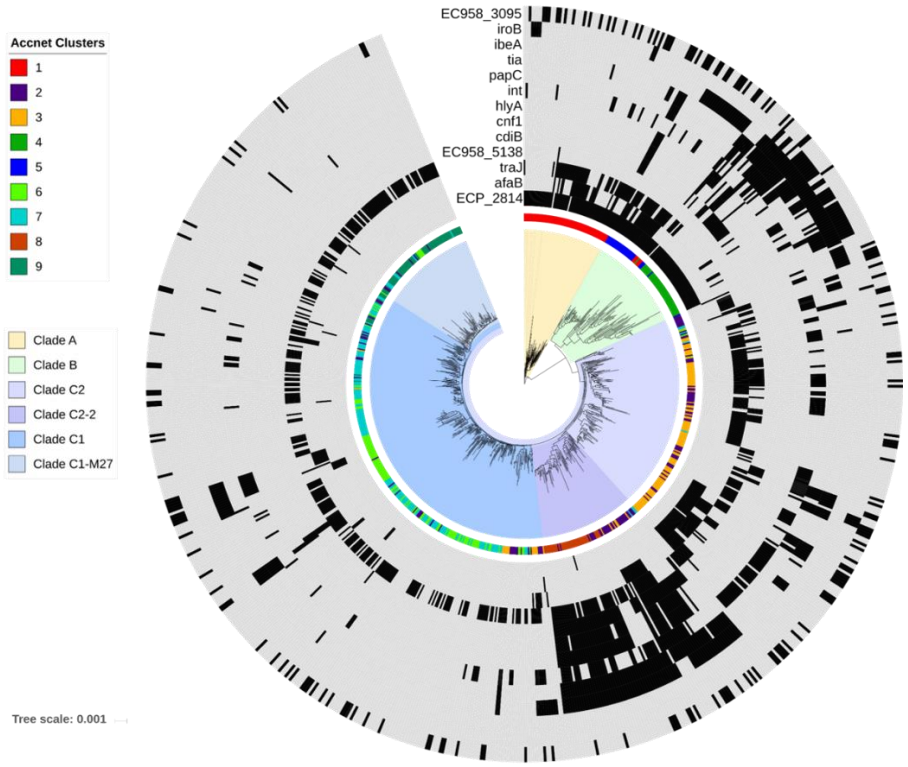


**Figure 11.** Virulence-related genes significantly associated to accessory genome of 797 NCBI STc131 samples (excluding those which occurred in fewer than 20 samples). Phylogenetic tree constructed by neighbor-joining method on ACCNET (accessory genome) distance matrix.

### 4.6.2 NCBI STc131 Virulence Profile

The NCBI STc131 assemblies contained significantly fewer ($p<0.01$) genes from the Virulence Factor Database (average of 180.8 virulence genes) than the non-STc131 phylogroup B2 NCBI dataset (average of 215.6).

A set of virulence genes/groups (*afaA*, *ibeA*, *iroN*, *traJ*, *cdtB*, *cnf1*, *hlyA*, *int*, *tia*, *papC*, and three hypothetical proteins: *ECP_2814*, *EC958_3095*, *EC598_5138*) were determined to be significant ($p<0.01$) within the accessory genome clusters of the NCBI STc131 using hypergeometric testing (Figure 11). The *ECP_2814* and *EC958_3095* genes are proposed to be type VI secretion system related, involved in association to host cell types[72–74], while EC598_5138 is proposed to be Cah calcium-binding autotransporter related, involved in autoaggregation and biofilm formation[75].

Virotype C is most frequent virulence profile (483 assemblies), making up the majority of Clade A, subclade C1, subclade C-M27, and a large proportion of subclade C2 (Figure 12). Of the 266 subclade C2 assemblies 42.1% are also virotype C, in contrast to the Ramón y Cajal sample in which only 14.3% of subclade C2 samples are virotype C. There are 110

assemblies with virulence profiles that do not fall within the Blanco virotype schema, 71 virotype D (55.4% of these virotype D3), 61 virotype A, 60 virotype E, and 8 virotype B. Of the subclade C2 accessory genome clusters, cluster 8 is closely associated with virotype E, while cluster 3 is associated with both virotypes A and C. Accnet cluster 2, which spans both subclades C1 and C2, is associated with all three Clade C virotypes: A, C, and E. The group of virotype E isolates with higher genomic similarity are reclassified here as subclade C2-2.

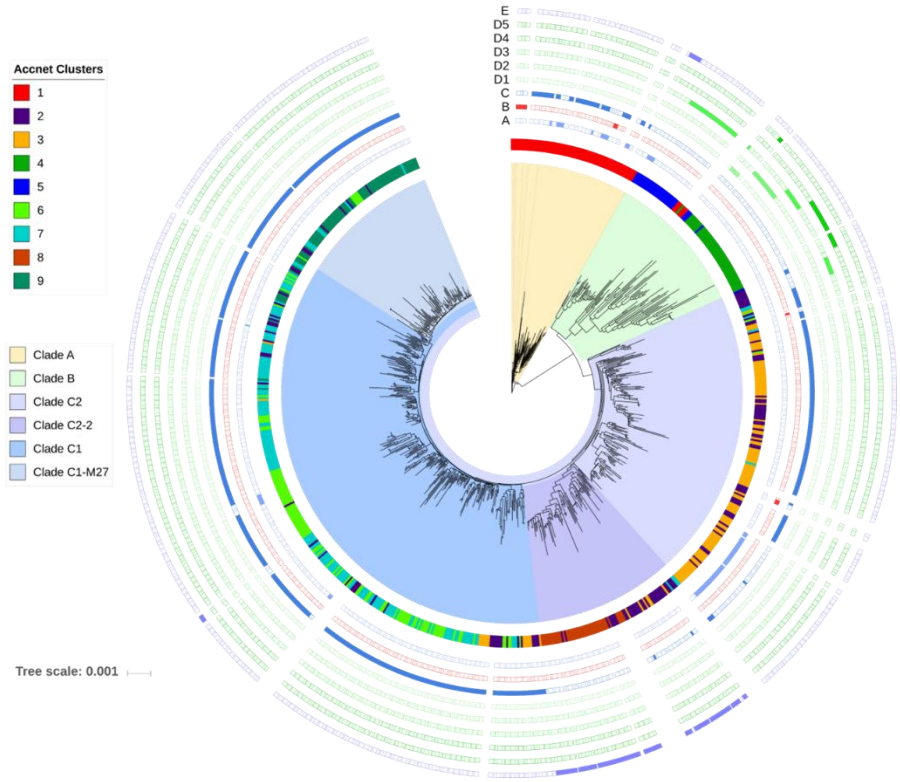


**Figure 12.** Neighbor joining tree from Mash distance matrix on 797 STc131 draft genomes curated from the NCBI database, Blanco virotypes compared against accessory genome clusters.

### 4.6.3 NCBI STc131 Antibiotic Resistance Genes

Antibiotic resistance genes within the NCBI STc131 sample were screened against the resfinder database. The NCBI STc131 assemblies contained significantly more ($p<0.01$) antibiotic resistance associated genes (average of 7.42 genes) than the non-STc131 phylogroup B2 dataset (average of 2.58). While 23.7% of the Ramón y Cajal STc131 sample and 47.2% of the NCBI STc131 assemblies are ESBL positive, only 2.8% of the non-STc131 phylogroup B2 dataset contained an ESBL.

Of the ESBL positive NCBI STc131, 83.4% of the 221 assemblies which contained *bla*$_{\text{CTX-M-15}}$ were within subclade C2, 79.6% of the 98 assemblies which contained *bla*$_{\text{CTX-M-27}}$ were within subclade C-M27, and 76.6% of the assemblies which contained *bla*$_{\text{CTX-M-14}}$ were within subclade C1. While 93.8% of the Ramón y Cajal *bla*$_{\text{CTX-M-15}}$ positive samples also

presented with the commonly-associated *bla*$_{OXA1}$ and *aac(6')-Ib-cr* genes, only 69.2% of the NCBI *bla*$_{CTX-M-15}$ positive samples contained both *bla*$_{OXA1}$ and *aac(6')-Ib-cr*.

In addition to the ESBL and ESBL-associated genes other antibiotic resistance genes were identified as significantly associated ($p<0.01$) to particular accessory genome clusters (*sul1*, *tet(a)*, *bla*$_{TEM-1B}$, *dfrA17, mph(A), aadA5, aph(6')-ld, aph(3")-lb*, *sul2*, *aac(3)-lla_1*, *tet(B)*, *aad12*, *dfrA12*, *dfrA14*, *bla*$_{KPC-2}$, *cmlA)* (Figure 13). Within these antibiotic resistance traits, clade B has the least robust antibiotic resistance profile and clade A and subclade C1 share a robust resistance profile comprised of *sul1, sul2, dfrA17, aadA5, mph(A), aph(6')-ld, aph(3")-lb*, *tet(A)* and *bla*$_{TEM-1B}$. The C-M27 subclade of C1 has a similar profile, with the *bla*$_{CTX-M-27}$ β-lactamase in the place of the *bla*$_{TEM-1B}$ β-lactamase. The *bla*$_{CTX-M-15}$ positive subclade C2 also generally lacks the *bla*$_{TEM-1B}$ β-lactamase, along with the *aph(6')-ld*, *aph(3")-lb*, and *sul2*.


**Figure 13.** Neighbor joining tree from Mash distance matrix on 797 STc131 draft genomes curated from the NCBI database, representing accessory genome clusters and significant antibiotic resistance traits (excluding those which occurred in fewer than 20 samples).

### 4.6.4 NCBI STc131 Replicon Sequence Type

The NCBI STc131 has a wide diversity of IncF replicon sequence types with 28 distinct FAB allele combinations. Some of these alleles differ by very few nucleotides, such as F36 and F105 which each differ from F31 by one point mutation. Here we see that the incF plasmid RSTs are largely clade and subclade specific: subclades C1 and C2 are each closely associated with ten distinct replicon sequence types, while Clades A and B each contain three to four unique RSTs specific to their clade in addition to two RSTs (F29:A-:B10, F4:A-:B10) that span between them (Figure 14). There are also instances of permeability which are generally unidirectional: there are smaller clusters subclade C1 isolates with RSTs largely associated with subclade C2 (F2:A1:B1, F31:A4:B1) and both clade A and subclade C2 contains clusters of isolates with RSTs largely associated with subclade C1 (F1:A2:B20 and F-:A2:B20, respectively).

While in the Hospital Ramón y Cajal samples, the subclade C2-2 isolates (those that had acquired the pathogenicity island PAI $II_{J96}$) were associated with incF plasmids of a distinct RST (F31/36:A4:B1), however within the NCBI STc131 samples the C2-2 subclade contains a wider variety of RSTs. These include the F31/36:A4:B1 RST, in addition to F2:A1:B- and F-:A2:B20.

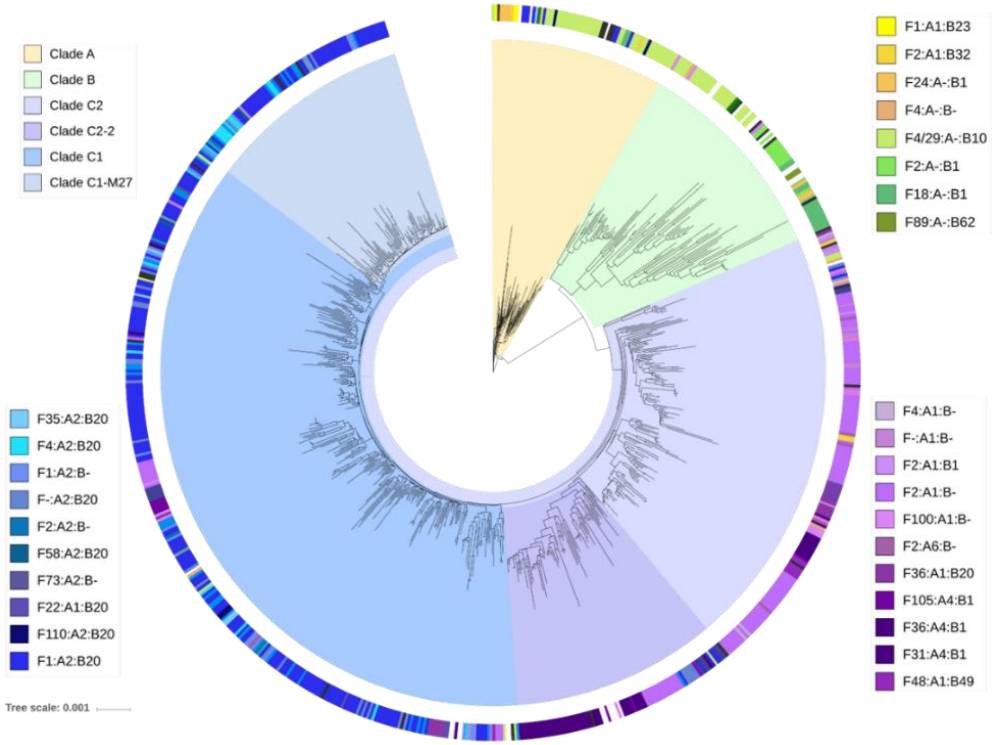


**Figure 14.** Neighbor joining tree from Mash distance matrix on 797 STc131 draft genomes curated from the NCBI database, representing IncF plasmid replicon allele-type, showing only first perfect hit (100% identity, 100% coverage) in cases where multiple copies of IncF FAB replicons were found.

# 5

# Conclusions and Future Directions

Using accent to delineate the accessory genome of the different Ramón y Cajal STc131 clades and subclades, as well as machine learning techniques such as clustering to analyze these relationships, reveals important information on the diversification within this clonal group. One such insight is the distinction of the clade C the accessory genome from the rest of STc131, with the highest proportion of genome products that are specific to the clusters identified within the clade. We see that the accessory genomes of clades B and C have a higher density of genome products related to replication, recombination and repair than clade A, which is consistent with the presence of elements related to MGE and increased accessory genome diversity, as well as many more products of unknown function. The accessory genome analysis also indicates that within STc131 there are cases where intra-clade variance rivals inter-clade variance: showing high separation between the accessory genome clusters within both clades B and C, as well as identifying accessory genome profiles that spanned multiple clades within the NCBI STc131 dataset.

The accessory genome clusters closely align with the distribution of the IncF plasmid RSTs, emphasizing the importance of IncF plasmids on the plasmidome of STc131 *E. coli* and on the accessory genome of this clonal group as a whole. Clades and subclades of STc131 are associated with specific plasmids and plasmid replicon sequence types, however there is varying degrees of flexibility and permeability between certain clades and for certain plasmids. Both the Ramón y Cajal sample and NCBI STc131 dataset show overlap in IncF replicon sequence type between subclades C1 and C2. This was reflected in the high similarity between F2:A1:B- plasmids identified within samples of both subclades, the composition of which differed primarily in the presence/absence of the CTX-M-15 antibiotic resistance cassette. The mosaicism of the incF plasmids is also displayed here, as many of the different RSTs were only distinguished by a change in one FAB replicon.

The identification of genes related to the epidemiology of STc131 and non-STc131 phylogroup B E. coli support previous understanding that STc131 has a higher frequency of antibiotic resistance traits and lower frequency of virulence-related genes compared to non-STc131 phylogroup B2[6,56]. However, the NCBI database of *E. coli* genomes is not necessarily representative of *E. coli* overall, with a bias towards highly studied strains. While the proportion of ESBL positive isolates collected in Hospital Ramón y Cajal from cases of bacteremia is also greater than in the non-STc131 phylogroup B2 dataset, it is almost half that of the NCBI STc131 dataset. The continued dominance of subclade C1 within the Ramón y Cajal STc131 sample through time, when it lacks the ESBL genes of subclade C2 as well as many of the virulence-related genes of clade B, highlights that other factors are important for the dissemination and persistence of STc131 subgroups.

All these data indicate the importance of the accessory genome, including the plasmid content of the isolates, in the dynamics of this STc131 population with each clade having a distinct plasmid, virulence, and antibiotic resistance signature. However, within the individual categories (plasmid content, virulence-related genes, antibiotic resistance genes) the boundaries between the clades are not absolute and we identify several points of overlap

for particular subclusters. These soft boundaries may contribute to the success of STc131 by balancing the specializations within the different clades with a degree of flexibility between them. Future steps include the reconstruction of a larger number set of plasmidomes to better delineate the relationships between clades, to identify the characteristics that determine which IncF plasmids are shared between clades or restricted to one, and the analysis of non-IncF plasmids such as Col plasmids which are associated with virulence potential of *E. coli*[76],

# Software Appendix

| Software | Version | Source |
|---|---|---|
| Python | 3.6.7 | Python programming language version 3.6.7 available at: https://www.python.org/ |
| Perl | v5.26.1 | Perl programming language version v5.26.1 available at: https://www.perl.org/ |
| RStudio | 3.4.4 | RStudio: Integrated Development for R v3.4.4 available at: http://www.rstudio.com/ |
| FastQC | 0.11.5 | FastQC: a quality control tool for high throughput sequence data v0.11.5 available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc |
| Trimmomatic | 0.38 | Bolger, A. M. *et al.* Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). |
| SPAdes | v3.13.0 | Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455–477 (2012). |
| Quast | v5.0.1 | Gurevich, A. *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013). |
| MLST | 2.15.2 | Torsten Seemann, mlst (2019) GitHub repository, https://github.com/tseemann/mlst |
| FimTyper | 1.0 | Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). |
| PointFinder | 2.0 | Zankari, E. *et al.* PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother* **72**, 2764–2768 (2017). |
| Snippy | 4.3.6 | Torsten Seemann, Snippy (2019) GitHub repository, https://github.com/tseemann/snippy |
| Gubbins | 2.3.1-1 | Croucher N. J. *et al.* "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". Nucleic Acids Research (2014). |
| IQtree | 1.6.1 | Nguyen, L.-T. *et al.* Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274 (2015). |
| Phylip | v3.696 | Chernomor, O. *et al.* Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol* **65**, 997–1008 (2016). |
| Mash | v2.1 | Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016). |
| Accnet | 3.0 | Val Lanza, Accnet, (2019), GitHub repository, https://github.com/valflanza/accnet |
| Eggnog Mapper | 2.0 | Eggnog Mapper (2019) Github repository, https://github.com/eggnogdb/eggnog-mapper |
| iTOL Interactive Tree Of Life web | V4 | Letunic, I. & Bork, P. Interactive Tree of Life (ITOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**, gkw290 (2016). |
| pMLST | 0.1.0 | Carattoli, A. *et al.* In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob Agents Chemother* **58**, 3895–3903 (2014). |

| | | |
|---|---|---|
| Mob-Suite | 1.4.9.1 | Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics* **4**, (2018). |
| ABRicate | 0.8.10 | Torsten Seemann, ABRicate (2019) GitHub repository, https://github.com/tseemann/abricate |
| Hyasp | Jan 2019 | Müller, R. & Chauve, C. HyAsP, a greedy tool for plasmids identification. *Bioinformatics* (2019). |
| NuRig | Beta | Val Lanza, NuRig, (2018), GitHub repository, https://github.com/valflanza/nurig |

# Glossary of Acronyms and Abbreviations

- **ABR:** Antibiotic Resistance
- **Ceph:** Cephalosporin
- **CGE:** Center for Genomic Epidemiology
- **cgMLST:** Core Genome Multi-Locus Sequence Type
- **COG:** Clusters of Orthologous Groups of proteins
- **EC:** Escherichia coli
- **ESBL:** Extended spectrum beta lactamase
- **ExPEC:** Extra-intestinal pathogenic Escherichia coli
- **FQ:** Fluoroquinolone
- **FQ-R:** Fluoroquinolone Resistant
- **FQ-S:** Fluoroquinolone Susceptible
- **GI:** Genomic Island
- **H30-R:** FimH30, FQ-R STc131. *Also referred to as subclade C1*
- **H30-Rx:** FimH30, ESBL positive, FQ-R STc131. *Also referred to as subclade C2*
- **H30-S:** FimH30 Type, FQ-S STc131. *Also referred to as Clade C0*
- **HPI:** Yersinia high-pathogenicity island
- **IncF:** Incompatibility Group F
- **MDR:** multidrug resistant
- **MGE:** Mobile Genetic Element
- **MLST:** Multi-Locus Sequence Type
- **NGS:** Next Generation Sequencing
- **PAI:** Pathogenicity island
- **pMLST:** Plasmid multi-locus sequence type
- **QRDR:** Quinolone-resistance determining region
- **Rep:** Replicon
- **RST:** Replicon Sequence Type
- **UTI:** Urinary tract infections

# Bibliography


1. Pitout, J. D. D. Extraintestinal pathogenic Escherichia coli: an update on antimicrobial resistance, laboratory diagnosis and treatment. *Expert Rev. Anti Infect. Ther.* **10**, 1165–1176 (2012).
2. Pitout, J. D. D. Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. *Front. Microbiol.* **3**, (2012).
3. WHO | Antimicrobial resistance: global report on surveillance 2014. *WHO* Available at: http://www.who.int/drugresistance/documents/surveillancereport/en/. (Accessed: 19th May 2019)
4. Day, M. J. *et al.* Population structure of Escherichia coli causing bacteraemia in the UK and Ireland between 2001 and 2010. *J. Antimicrob. Chemother.* **71**, 2139–2142 (2016).
5. Peirano, G. *et al.* Global Incidence of Carbapenemase-Producing Escherichia coli ST131. *Emerg. Infect. Dis.* **20**, 1928–1931 (2014).
6. Nicolas-Chanoine, M.-H. *et al.* Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *J. Antimicrob. Chemother.* **61**, 273–281 (2008).
7. Platell, J. L., Johnson, J. R., Cobbold, R. N. & Trott, D. J. Multidrug-resistant extraintestinal pathogenic Escherichia coli of sequence type ST131 in animals and foods. *Vet. Microbiol.* **153**, 99–108 (2011).
8. Johnson, J. R. *et al.* Comparison of Escherichia coli ST131 pulsotypes, by epidemiologic traits, 1967-2009. *Emerg. Infect. Dis.* **18**, 598–607 (2012).
9. Blanco, J. *et al.* Four Main Virotypes among Extended-Spectrum-β-Lactamase-Producing Isolates of Escherichia coli O25b:H4-B2-ST131: Bacterial, Epidemiological, and Clinical Characteristics. *J. Clin. Microbiol.* **51**, 3358–3367 (2013).
10. Novais, Â. *et al.* Characterization of Globally Spread Escherichia coli ST131 Isolates (1991 to 2010). *Antimicrob. Agents Chemother.* **56**, 3973–3976 (2012).
11. Zakour, N. L. B. *et al.* Sequential Acquisition of Virulence and Fluoroquinolone Resistance Has Shaped the Evolution of Escherichia coli ST131. *mBio* **7**, e00347-16 (2016).
12. Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. *mBio* **7**, (2016).
13. Mathers, A. J., Peirano, G. & Pitout, J. D. D. The Role of Epidemic Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-Resistant Enterobacteriaceae. *Clin. Microbiol. Rev.* **28**, 565 (2015).
14. Lanza, V. F. *et al.* Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. *PLOS Genet.* **10**, e1004766 (2014).
15. Stoesser, N. *et al.* Complete Sequencing of Plasmids Containing blaOXA-163 and blaOXA-48 in Escherichia coli Sequence Type 131. *Antimicrob. Agents Chemother.* **60**, 6948–6951 (2016).
16. Lanza, V. F., Baquero, F., de la Cruz, F. & Coque, T. M. AcCNET (Accessory

Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. *Bioinformatics* **33**, 283–285 (2017).
17. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **5**, 58–65 (2013).
18. Boyd, E. F. & Hartl, D. L. Chromosomal Regions Specific to Pathogenic Isolates of Escherichia coli Have a Phylogenetically Clustered Distribution. *J. Bacteriol.* **180**, 1159–1165 (1998).
19. Coque, T. M. *et al.* Dissemination of Clonally Related Escherichia coli Strains Expressing Extended-Spectrum β-Lactamase CTX-M-15. *Emerg. Infect. Dis.* **14**, 195–200 (2008).
20. Blanco, J. *et al.* National survey of Escherichia coli causing extraintestinal infections reveals the spread of drug-resistant clonal groups O25b:H4-B2-ST131, O15:H1-D-ST393 and CGA-D-ST69 with high virulence gene content in Spain. *J. Antimicrob. Chemother.* **66**, 2011–2021 (2011).
21. Le Gall, T. *et al.* Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. *Mol. Biol. Evol.* **24**, 2373–2384 (2007).
22. Johnson, J. R., Porter, S., Thuras, P. & Castanheira, M. The Pandemic H30 Subclone of Sequence Type 131 (ST131) as the Leading Cause of Multidrug-Resistant Escherichia coli Infections in the United States (2011–2012). *Open Forum Infect. Dis.* **4**, (2017).
23. Petty, N. K. *et al.* Global dissemination of a multidrug resistant Escherichia coli clone. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5694–5699 (2014).
24. Clermont, O. *et al.* Rapid detection of the O25b-ST131 clone of Escherichia coli encompassing the CTX-M-15-producing strains. *J. Antimicrob. Chemother.* **64**, 274–277 (2009).
25. Nicolas-Chanoine, M.-H., Bertrand, X. & Madec, J.-Y. Escherichia coli ST131, an Intriguing Clonal Group. *Clin. Microbiol. Rev.* **27**, 543–574 (2014).
26. Weissman, S. J. *et al.* High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli. *Appl. Environ. Microbiol.* **78**, 1353–1360 (2012).
27. Johnson, J. R. *et al.* Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. *J. Infect. Dis.* **207**, 919–928 (2013).
28. Dias, R. C. S., Moreira, B. M. & Riley, L. W. Use of fimH single-nucleotide polymorphisms for strain typing of clinical isolates of Escherichia coli for epidemiologic investigation. *J. Clin. Microbiol.* **48**, 483–488 (2010).
29. Tartof, S. Y., Solberg, O. D. & Riley, L. W. Genotypic analyses of uropathogenic Escherichia coli based on fimH single nucleotide polymorphisms (SNPs). *J. Med. Microbiol.* **56**, 1363–1369 (2007).
30. Downing, T. Tackling Drug Resistant Infection Outbreaks of Global Pandemic Escherichia coli ST131 Using Evolutionary and Epidemiological Genomics. *Microorganisms* **3**, 236–267 (2015).
31. Peirano, G. *et al.* Characteristics of Escherichia coli Sequence Type 131 Isolates That Produce Extended-Spectrum β-Lactamases: Global Distribution of the H30-Rx

Sublineage. *Antimicrob. Agents Chemother.* **58**, 3762–3767 (2014).
32. Gaiarsa, S. *et al.* Comparative Analysis of the Two Acinetobacter baumannii Multilocus Sequence Typing (MLST) Schemes. *Front. Microbiol.* **10**, (2019).
33. Ortega, A. *et al.* Carbapenemase-producing Escherichia coli is becoming more prevalent in Spain mainly because of the polyclonal dissemination of OXA-48. *J. Antimicrob. Chemother.* **71**, 2131–2138 (2016).
34. Johnson, J. R. *et al.* Rapid and Specific Detection, Molecular Epidemiology, and Experimental Virulence of the O16 Subgroup within Escherichia coli Sequence Type 131. *J. Clin. Microbiol.* **52**, 1358–1365 (2014).
35. Dahbi, G. *et al.* Emergence of new variants of ST131 clonal group among extraintestinal pathogenic Escherichia coli producing extended-spectrum β-lactamases. *Int. J. Antimicrob. Agents* **42**, 347–351 (2013).
36. Barnard, F. M. & Maxwell, A. Interaction between DNA Gyrase and Quinolones: Effects of Alanine Mutations at GyrA Subunit Residues Ser83 and Asp87. *Antimicrob. Agents Chemother.* **45**, 1994–2000 (2001).
37. Yoshida, H., Bogaki, M., Nakamura, M. & Nakamura, S. Quinolone resistance-determining region in the DNA gyrase gyrA gene of Escherichia coli. *Antimicrob. Agents Chemother.* **34**, 1271–1272 (1990).
38. Pitout, J. D. D. & DeVinney, R. Escherichia coli ST131: a multidrug-resistant clone primed for global domination. *F1000Research* **6**, 195 (2017).
39. Price, L. B. *et al.* The Epidemic of Extended-Spectrum-β-Lactamase-Producing Escherichia coli ST131 Is Driven by a Single Highly Pathogenic Subclone, H30-Rx. *mBio* **4**, (2013).
40. Peirano, G. & Pitout, J. D. D. Fluoroquinolone-resistant Escherichia coli sequence type 131 isolates causing bloodstream infections in a canadian region with a centralized laboratory system: rapid emergence of the H30-Rx sublineage. *Antimicrob. Agents Chemother.* **58**, 2699–2703 (2014).
41. Bauernfeind, A., Grimm, H. & Schweighart, S. A new plasmidic cefotaximase in a clinical isolate of Escherichia coli. *Infection* **18**, 294–298 (1990).
42. Cantón, R., González-Alba, J. M. & Galán, J. C. CTX-M Enzymes: Origin and Diffusion. *Front. Microbiol.* **3**, (2012).
43. Adamski, C. J. *et al.* Molecular basis for the catalytic specificity of the CTX-M extended-spectrum β-lactamases. *Biochemistry* **54**, 447–457 (2015).
44. Tian, G.-B. *et al.* CTX-M-137, a hybrid of CTX-M-14-like and CTX-M-15-like β-lactamases identified in an Escherichia coli clinical isolate. *J. Antimicrob. Chemother.* **69**, 2081–2085 (2014).
45. Matsumura, Y. *et al.* CTX-M-27- and CTX-M-14-producing, ciprofloxacin-resistant Escherichia coli of the H30 subclonal group within ST131 drive a Japanese regional ESBL epidemic. *J. Antimicrob. Chemother.* **70**, 1639–1649 (2015).
46. Matsumura, Y. *et al.* Global Escherichia coli Sequence Type 131 Clade with blaCTX-M-27 Gene. *Emerg. Infect. Dis.* **22**, 1900–1907 (2016).
47. Emami, S., Shafiee, A. & Foroumadi, A. Quinolones: Recent Structural and Clinical Developments. *Iran. J. Pharm. Res.* **0**, 123–136 (2010).
48. Coque, T. M., Baquero, F. & Canton, R. Increasing prevalence of ESBL-producing

Enterobacteriaceae in Europe. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **13**, (2008).
49. Brisse, S. *et al.* Phylogenetic Distribution of CTX-M- and Non-Extended-Spectrum-β-Lactamase-Producing Escherichia coli Isolates: Group B2 Isolates, Except Clone ST131, Rarely Produce CTX-M Enzymes. *J. Clin. Microbiol.* **50**, 2974–2981 (2012).
50. Bonnet, R. Growing Group of Extended-Spectrum β-Lactamases: the CTX-M Enzymes. *Antimicrob. Agents Chemother.* **48**, 1–14 (2004).
51. Johnson, J. R. *et al.* Escherichia coli Sequence Type 131 H30 Is the Main Driver of Emerging Extended-Spectrum-β-Lactamase-Producing E. coli at a Tertiary Care Center. *mSphere* **1**, (2016).
52. Olesen, B. *et al.* Prevalence and Characteristics of the Epidemic Multiresistant Escherichia coli ST131 Clonal Group among Extended-Spectrum Beta-Lactamase-Producing E. coli Isolates in Copenhagen, Denmark. *J. Clin. Microbiol.* **51**, 1779–1785 (2013).
53. Johnson, J. R. *et al.* Association of carboxylesterase B electrophoretic pattern with presence and expression of urovirulence factor determinants and antimicrobial resistance among strains of Escherichia coli that cause urosepsis. *Infect. Immun.* **59**, 2311–2315 (1991).
54. Vimont, S. *et al.* The CTX-M-15-Producing Escherichia coli Clone O25b: H4-ST131 Has High Intestine Colonization and Urinary Tract Infection Abilities. *PLoS ONE* **7**, (2012).
55. Clermont, O. *et al.* The CTX-M-15-producing Escherichia coli diffusing clone belongs to a highly virulent B2 phylogenetic subgroup. *J. Antimicrob. Chemother.* **61**, 1024–1028 (2008).
56. Lavigne, J.-P. *et al.* Virulence Potential and Genomic Mapping of the Worldwide Clone Escherichia coli ST131. *PLoS ONE* **7**, (2012).
57. Johnson, J. R., Porter, S. B., Zhanel, G., Kuskowski, M. A. & Denamur, E. Virulence of Escherichia coli Clinical Isolates in a Murine Sepsis Model in Relation to Sequence Type ST131 Status, Fluoroquinolone Resistance, and Virulence Genotype. *Infect. Immun.* **80**, 1554–1562 (2012).
58. Mora, A. *et al.* Virulence Patterns in a Murine Sepsis Model of ST131 Escherichia coli Clinical Isolates Belonging to Serotypes O25b:H4 and O16:H5 Are Associated to Specific Virotypes. *PLOS ONE* **9**, e87025 (2014).
59. Schubert, S., Picard, B., Gouriou, S., Heesemann, J. & Denamur, E. Yersinia High-Pathogenicity Island Contributes to Virulence in Escherichia coli Causing Extraintestinal Infections. *Infect. Immun.* **70**, 5335–5337 (2002).
60. Massot, M. *et al.* Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of Escherichia coli from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology* **162**, 642–650 (2016).
61. Hacker, J. & Kaper, J. B. *Pathogenicity Islands and the Evolution of Pathogenic Microbes*. (Springer Science & Business Media, 2012).
62. Zong, Z. Complete sequence of pJIE186-2, a plasmid carrying multiple virulence factors from a sequence type 131 Escherichia coli O25 strain. *Antimicrob. Agents Chemother.* **57**, 597–600 (2013).

63. Johnson, T. J. *et al.* Separate F-Type Plasmids Have Shaped the Evolution of the H30 Subclone of Escherichia coli Sequence Type 131. *mSphere* **1**, e00121-16 (2016).
64. Matsumura, Y. *et al.* Association of Fluoroquinolone Resistance, Virulence Genes, and IncF Plasmids with Extended-Spectrum-β-Lactamase-Producing Escherichia coli Sequence Type 131 (ST131) and ST405 Clonal Groups. *Antimicrob. Agents Chemother.* **57**, 4736–4742 (2013).
65. Johnson, T. J. *et al.* Plasmid Replicon Typing of Commensal and Pathogenic Escherichia coli Isolates. *Appl. Environ. Microbiol.* **73**, 1976–1983 (2007).
66. Villa, L., García-Fernández, A., Fortini, D. & Carattoli, A. Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *J. Antimicrob. Chemother.* **65**, 2518–2529 (2010).
67. Rafaï, C. *et al.* Dissemination of IncF-type plasmids in multiresistant CTX-M-15-producing Enterobacteriaceae isolates from surgical-site infections in Bangui, Central African Republic. *BMC Microbiol.* **15**, (2015).
68. Phan, M. D. *et al.* Molecular Characterization of a Multidrug Resistance IncF Plasmid from the Globally Disseminated Escherichia coli ST131 Clone. *PLOS ONE* **10**, e0122369 (2015).
69. Yang, Q. E. *et al.* IncF plasmid diversity in multi-drug resistant Escherichia coli strains from animals in China. *Front. Microbiol.* **6**, (2015).
70. Clermont, O., Bonacorsi, S. & Bingen, E. Rapid and Simple Determination of the Escherichia coli Phylogenetic Group. *Appl. Environ. Microbiol.* **66**, 4555–4558 (2000).
71. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
72. Aschtgen, M.-S., Gavioli, M., Dessen, A., Lloubès, R. & Cascales, E. The SciZ protein anchors the enteroaggregative Escherichia coli Type VI secretion system to the cell wall. *Mol. Microbiol.* **75**, 886–899 (2010).
73. Aschtgen, M.-S., Bernard, C. S., De Bentzmann, S., Lloubès, R. & Cascales, E. SciN Is an Outer Membrane Lipoprotein Required for Type VI Secretion in Enteroaggregative Escherichia coli. *J. Bacteriol.* **190**, 7523–7531 (2008).
74. Aschtgen, M.-S., Zoued, A., Lloubès, R., Journet, L. & Cascales, E. The C-tail anchored TssL subunit, an essential protein of the enteroaggregative Escherichia coli Sci-1 Type VI secretion system, is inserted by YidC. *MicrobiologyOpen* **1**, 71–82 (2012).
75. Torres, A. G. *et al.* Characterization of Cah, a calcium-binding and heat-extractable autotransporter protein of enterohaemorrhagic Escherichia coli. *Mol. Microbiol.* **45**, 951–966 (2002).
76. Milch, H., Nikolnikov, S. & Czirók, E. Escherichia coli Col V plasmids and their role in pathogenicity. *Acta Microbiol. Hung.* **31**, 117–125 (1984).